AFRL-RY-WP-TR-2018-0030

A STUDY OF THE CHARGE TRAP TRANSISTOR (CTT) FOR POST-FAB MODIFICATION OF WAFERS

Subramanian S. Iyer

University of California Los Angeles

APRIL 2018
Final Report

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals.

Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RY-WP-TR-2018-0030 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

// Signature//

CHRISTOPHER A. BOZADA, Program Manager
Aerospace Components & Subsystems Division

// Signature//

JAMES M. SATTLER, Lt Col, USAF
Deputy Chief
Aerospace Components & Subsystems Division
Sensors Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

*Disseminated copies will show "//Signature//" stamped or typed above the signature blocks.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
April 2018	Final	13 June 2016 – 13 December 2017

4. TITLE AND SUBTITLE
A STUDY OF THE CHARGE TRAP TRANSISTOR (CTT) FOR POST-FAB MODIFICATION OF WAFERS

5a. CONTRACT NUMBER
FA8650-16-1-7648

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)
Subramanian S. Iyer

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER
N/A

5f. WORK UNIT NUMBER
Y1CA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California Los Angeles
11000 Kinross Avenue, Suite 102
Los Angeles, CA 90095-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Sensors Directorate
Wright-Patterson Air Force Base, OH 45433-7320
Air Force Materiel Command
United States Air Force

Defense Advanced
Research Projects Agency
DARPA/MTO
675 North Randolph Street
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RYD

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RY-WP-TR-2018-0030

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This material is based on research sponsored by Air Force Research laboratory (AFRL) and the Defense Advanced Research Agency (DARPA) under agreement number FA8650-16-1-7648. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation herein. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies of endorsements, either expressed or implied, of AFRL and the DARPA or the U.S. Government. Report contains color.

14. ABSTRACT
This report summarizes a year-long study on the applicability of the charge trap transistor (CTT) for embedded memory applications. Two case uses are considered (1) as a digital multi-time programmable memory and (2) as a reprogrammable analog memory. Experimental data reveals that a CTT for analog memory applications possesses promising characteristics for implementing synapses in neural networks, such as spike-timing dependent plasticity, very fine tunability, weight-dependent plasticity, and low power consumption. Ongoing efforts include the design and tapeout of a CTT-based neuromorphic chip for digit recognition, and more elaborate designs that address programming time, scalability, power/area reduction, redundancy, unsupervised learning, etc. in the long-term.

15. SUBJECT TERMS
charge trap transistor, digital multi-time programmable memory, analog memory

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Christopher Bozada
Unclassified	Unclassified	Unclassified	SAR	51	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

Table of Contents

Section **Page**

List of Figures

Figure **Page**

Figure **Page**

Figure **Page**

List of Tables

Table **Page**

Abstract

This report summarizes a year-long study sponsored by the Defense Advanced Research Projects Agency (DARPA) on the applicability of the Charge Trap Transistor (CTT) for embedded memory applications. Two case uses are considered (1) as a digital Multi-Time Programmable Memory (MTPM) and (2) as a reprogrammable analog memory. High-K materials, used as gate dielectrics in 45nm (and below) node, exhibit significant charge trapping. The trapped charge may be leveraged to dynamically change the threshold of a Metal-Oxide-Semiconductor (MOS) transistor by electrical means. We have explored the use of CTTs for both digital and analog memory and describe the results in this report.

The charge trapping effect is greatly enhanced by self-heating, the layout geometry of the CTT cell, therefore, needs to be optimized to get the maximum change in threshold voltage (V_T). Self-heating is primarily achieved by driving the device hard with both gate voltage as well as drain to source voltage. The stability of the threshold voltage change is dependent on the emission cross-section of the traps. Deeper traps are more stable. A further effect is the presence of a thin interfacial oxide, thickness of which affects the trapping characteristics and effective threshold voltage perturbation. Dielectrics with significant trapped charge exhibit two issues: increased gate leakage, presumably due to trap assisted tunneling, and a potential for lower gate reliability, *e.g.* Time Dependent Dielectric Breakdown (TDDB). Both issues are exacerbated with reduced quality of the dielectric. The issues can be mitigated by applying innovative circuit design techniques. While the CTT effect is present in Complementary Metal-Oxide-Semiconductor (CMOS) technologies that employ Hi K dielectrics, the process architecture affects the trapping characteristics and the implications for the circuit design (including approaches to mitigate gate induced drain leakage (GIDL), lower self-heating in Bulk CMOS) and TDDB) are summarized in Table 1. The charge retention at operating temperature of 85C is estimated to be >10 years.

Table 1. CTT Evolution as a Function of Node and Technology

Node/Technology	Process Architectural Features	Impact on CTT Device	Design implications
32nm PDSOI	Planar SOI, Hi K gate 1st Tox= 1.2nm	Cell: 0.109mm2 VT shift max ~300mV (Tgt: 150mV) (@Ig <10μA) Erase: ~70% of Vt max	Overwrite protection for target VT shift. Cell initialization in manufacturing (MTPM). 2~4x OTPM cell for 70% erase (MTPM).
22nm PDSOI	Planar SOI, Hi K gate 1st Tox= 1.0nm	Cell = 0.144mm2 VT shift max ~300mV (tgt: 150mV) Erase: ~70% of Vt max	Overwrite protection for target VT shift. Cell initialization in manufacturing (MTPM). 2~4x OTPM cell for 70% erase (MTPM).
22nm FDSOI	Planar SOI, Hi K gate 1st, FD Tox=1.25nm	Cell: (NA) VT shift max ~300mV (tgt. 150mV) Erase: > 90% (1st erase) of Vt max	Overwrite protection for target VT shift. Cell initialization in manufacturing (MTPM). ~2x OTPM cell for >90% erase (MTPM).
14nm PD SOI FinFET	FinFET on SOI ,Hi K, gate last (lower Temp) Tox= 1.2nm	Cell: 0.121 (12FINs x 1) VT shift max ~100mV (Tgt: 50mV) Erase: ~50% (more trap assisted Leakage, less recovery, more g_m degrade)	Overwrite protection for target VT shift Cell initialization in manufacturing (MTPM). 4~8X OTPM cell for 50% erase (MTPM).
14nm Bulk	Bulk FinFET, Hi K, gate last (lower Temp) Tox= 1.33nm Less Self Heating (bulk)	Cell: 0.1411mm2 (6FINS x 2) VT shift max ~100mV (tgt: 50mV) Erase: ~50% (more trap assisted Leakage, less recovery, more g_m degrade) GIDL (bulk Substrate)	Overwrite protection for target VT shift. Cell initialization in manufacturing (MTPM). 4~8X (OTPM cells) for 50% erase (MTPM). Triple well for GIDL reduction (MTPM). Over erase protection (MTPM). Self-heating design assist (MTPM).

Another part of our focus has been to explore alternative memory modalities – especially analog memory using the CTT. We have studied the characteristics of CTT for analog memory, including very fine tuning of the channel conductance obtained by small changes in the threshold voltage and measuring channel conductance in the subthreshold region. Using this approach we have been able to demonstrate spike-timing dependent plasticity, and weight-dependent plasticity. A proof-of-concept unsupervised learning network was demonstrated to cluster stylized letters and noisy versions of them. It is found that the network is very robust to device variations. In addition, a CTT-based inference engine is being investigated for compact, ultra-low-power neuromorphic systems. We have been advised that DARPA will support a follow-on project that aims to implement the CTT in a neuromorphic chip. We will no doubt become more aware of nuances of the CTT in that work and will keep DARPA posted with regular updates.

1. Physics of CTT for Digital Memory

1.1 Introduction

Oxygen vacancy related traps and charge trapping are well-known phenomena in HfO_2, a commonly used gate dielectric material in high-k-metal-gate (HKMG) CMOS technologies [1][2][3]. It is also known that bias stress induced charge trapping and defect generation in HfO_2 are strongly accelerated by temperature [4][5]. While charge trapping is typically considered to be a nuisance, as it is a source of variability in devices and in turn circuits, we propose and demonstrate that this propensity for charge trapping in HfO_2 can be utilized as a feature for embedded non-volatile memory (NVM) applications in HKMG CMOS technologies. Charge trapping in high-k dielectrics such as HfO_2 for non-volatile memory applications has been proposed before [2][3]. In this work, however, we study and demonstrate how charge trapping in HfO_2 can be exploited for achieving a multiple-time programmable, fully HKMG CMOS process compatible, non-volatile memory with excellent retention characteristics and logic-compatible low-power operation, using only standard, as fabricated CMOS logic transistors. We demonstrate how the use of HKMG logic field-effect transistors (FETs) under modified operation conditions can result in enhanced charge trapping in the gate dielectric material, resulting in a threshold voltage shift (ΔV_T), which can in turn be utilized as a non-volatile data storage mechanism. The fundamentals of enhanced charge trapping and enhanced charge stability under the proposed operation conditions are studied. In addition, operation principles of the memory devices, scaling, reliability, and modeling of the device behavior are discussed. Furthermore, fully-functional product prototype memory arrays integrated on 32nm SOI, 22nm SOI, 14nm SOI, and 14nm bulk FinFET technologies are demonstrated.

1.2 Overview of Charge Trap Transistors

The fundamental principle of operation behind CTT based NVM is modulation of the device V_T by charge trapped in the high-k dielectric, where each unique V_T value can be interpreted as a unique bit *e.g.*, "0" and "1" for two unique V_T levels. The V_T of the device is governed by the following equation, where Q_{ox} is the quantity that is modulated due to the charge trapped in the gate dielectric

$$V_T = -\frac{Q_{dm}}{C_{ox}} + 2\Phi_F + V_{FB} - \frac{\boldsymbol{Q_{ox}}}{C_{ox}} \quad .$$

A schematic depicting the basic operation of a CTT memory device is shown in Figure 1. Note that this effect is equally applicable to planar FET as well as FinFET based devices, as will be demonstrated in the following discussions.

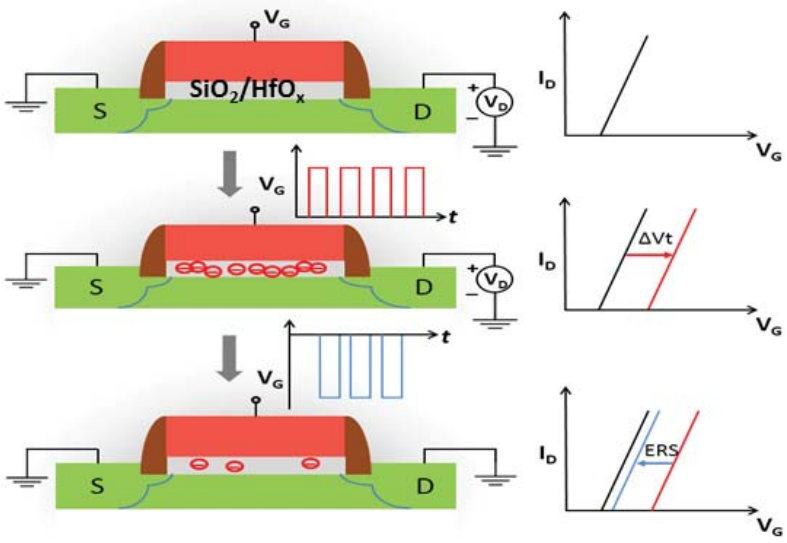


Figure 1: Schematic depicting the Basic Operation of a CTT Memory Device
(Equally applicable to planar FET as well as FinFET based CTTs)

To understand the dynamic charge trapping behavior of the device at various bias conditions, ΔV_T of devices fabricated using 22nm SOI technology [6] are measured during pulsed gate voltage ramp sweeps (PVRS) for various fixed drain bias (V_d) conditions. Details on the PVRS technique can be found in [7]. In this study, gate bias (V_g) is applied using 10ms pulses of increasing magnitudes in increments 50mV. After each pulse, V_T of the device is measured within ~10ms. Each device is ramped until breakdown; the measured ΔV_T values prior to breakdown are shown in Figure 2. At higher V_d (higher lateral field and increased self-heating), equivalent ΔV_T values are achievable at substantially lower V_g. This is attributed to the impact of an enhanced level of hot carrier injection and charge trapping with increasing V_d, as well as to enhanced charge trapping due to device self-heating [8] with increasing V_d. In addition, the maximum achievable ΔV_T prior to device breakdown initially increases and then starts to decrease with increasing V_d. The breakdown of devices under low V_d conditions is electric field driven (high gate-to-drain bias, V_{gd}), whereas the breakdown of devices under high V_d conditions (which happens at much lower V_{gd}) is self-heating driven, which is a well-known phenomenon [9]. Shifts in ΔV_T vs. V_g trends before hard breakdown may be indicative of the beginning of soft breakdown [10].

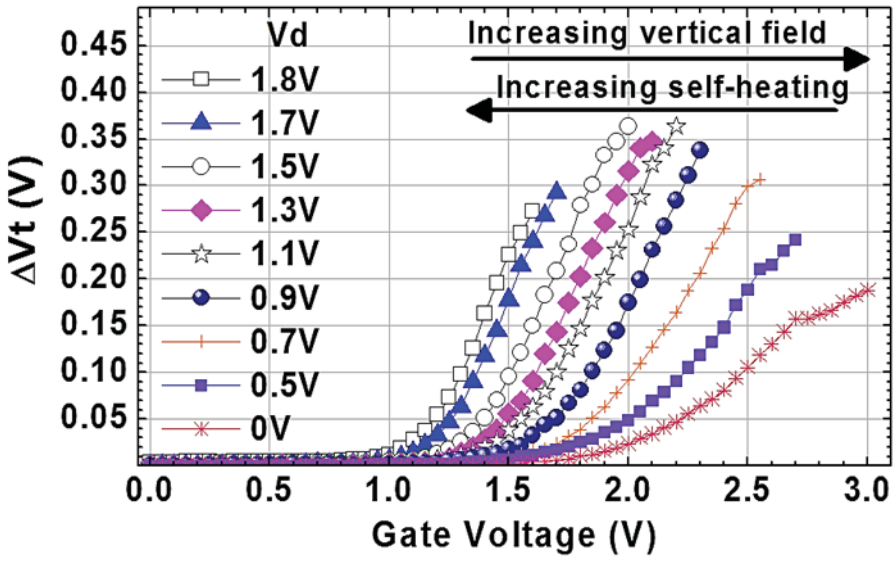


Figure 2: ΔV_T as a Function of PVRS Stress with 10ms Pulses at various fixed V_d
(W_{ch}=1.04µm, L=20nm)

To study the impact of device scaling, identical stress pulses (35ms at V_g=2V and V_d=1.3V) are applied to devices with the same channel width (W_{ch}) and different channel length (L). V_T of the devices is measured within 10ms. It can be observed that ΔV_T increases as L decreases (shown in Figure 3(a)), which is expected and consistent with increasing levels of hot carriers and self-heating (due to increase in lateral field), and decreasing V_T (due to short-channel effects) with decreasing L. Our results however, also show that when identical stress pulses are applied to devices with the same L and different W_{ch}, ΔV_T increases with W_{ch} (shown in Figure 3(b)). This phenomenon of ΔV_T varying with W_{ch} (even when vertical and lateral fields and L are the same) is not readily explained by a uniform injection mechanism along W_{ch} and is attributed to the impact of self-heating, which is strongly modulated by W_{ch}. In this work, the chuck temperature of probe station was kept at 25C unless otherwise stated.

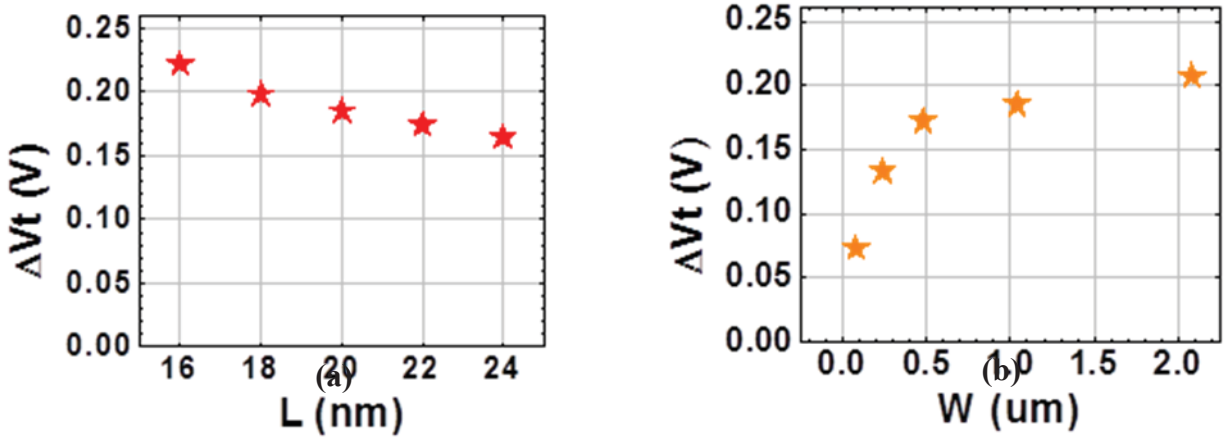


Figure 3: Measured ΔV_T vs. (a) Device L=1.04µm and (b) Device W_{ch}=20nm

To better understand and quantify this phenomenon and to separate the impact of electric field from the thermal effects, single-finger devices vs. multi-finger (split-channel) devices are studied. Both devices have a total W_{ch} of 1.04µm. Each channel within multi-finger devices is separated by trench isolation and has a width of $W_{ch}/4$. Both devices are identical to each other except for the channel width and have a channel length of 20nm. Layout of the two devices is shown in, respectively, Figure 4(a) and Figure 4(b). First, thermal profiles of the channel of the two devices are analyzed. Thermal simulations were carried out using finite element analysis (Comsol™). Full 3D structural simulations of the devices are analyzed and solved for temperature distribution and heat flux. Channel temperature (T) profile of the two devices for applied power density of 4mW/µm is shown in Figure 5(a) and Figure 5(b). It can be seen that W_{ch}/active area significantly modulates device self-heating. In multi-finger devices, the area for vertical heat flow is effectively larger than the area for power dissipation. In addition, the larger area for lateral heat dissipation and the higher number of contacts per unit width in multi-finger devices are lead to higher heat dissipation and thus lower thermal resistance (R_{th}), as compared to single-finger devices. The extracted R_{th} for single-finger devices is found to be ~1.3x compared to multi-finger devices. From simulations, the devices reach thermal equilibrium within several hundred nanoseconds, as shown in Figure 6.

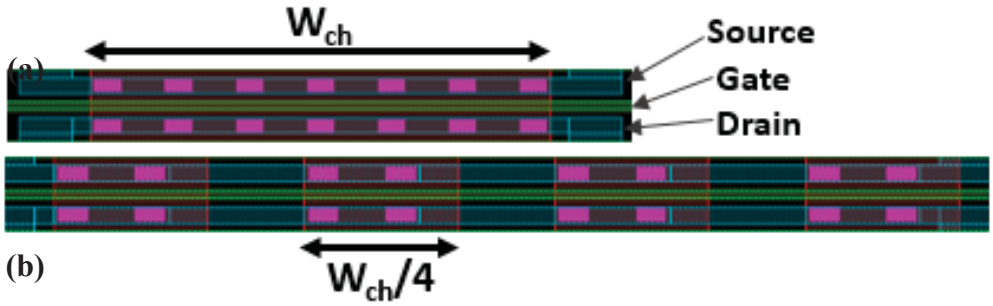


Figure 4: (a) Single-Finger and (b) Multi-Finger Device Layout

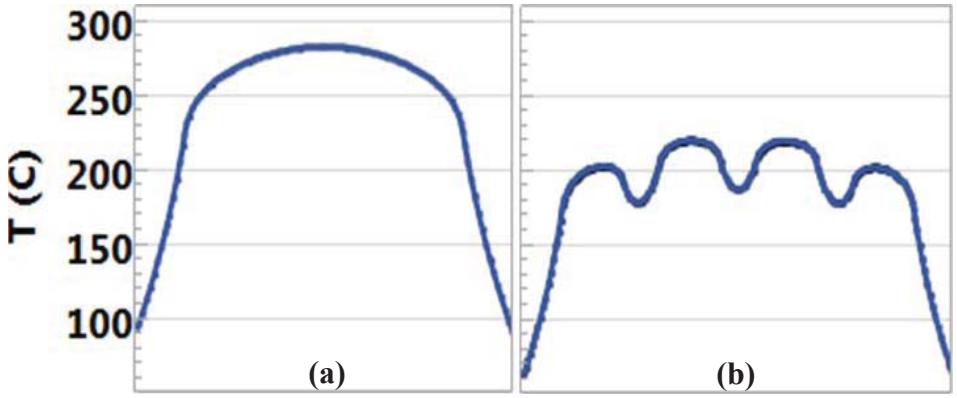


Figure 5: Steady-State Thermal Profiles for (a) Single-Channel and (b) Multi-Channel Device in the W_{ch} Direction, for an applied Power of 4mW/µm

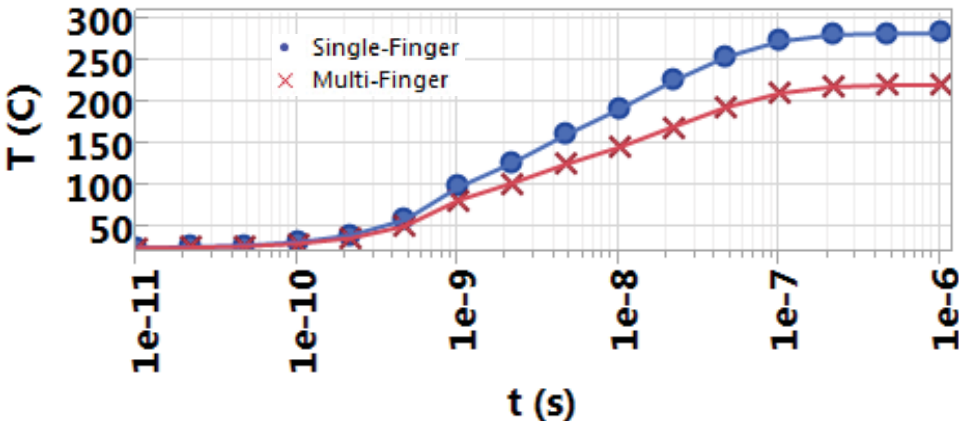


Figure 6: Rise in Channel Temperature vs. Stress Time
(4mW/μm applied power)

Measured ΔV_T vs. applied power density for the same devices that were used in the thermal simulations is shown in Figure 7(a). The power is varied by changing V_d while V_g=2V. It can be seen that for the same power density, ΔV_T for the single-channel device is considerably higher as compared to the split-channel device, and the difference is greater at higher power densities. However, when plotted as a function of the calculated channel temperature (shown in Figure 7(b)), the characteristics of ΔV_T for the two devices are almost identical, except at very high temperatures where the single-channel device seems to have slightly higher ΔV_T. In other words, ΔV_T behavior of the devices exhibits a significantly stronger correlation to the device self-heating temperature than to the applied power density. It is concluded from these results that the device self-heating temperature is a significant factor in modulating the charge trapping behavior.

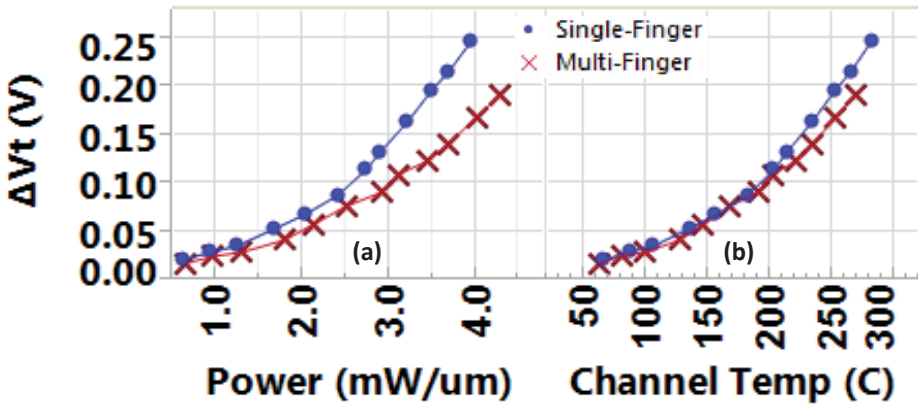


Figure 7: Measured ΔV_T vs. (a) Applied Power Density and (b) Channel T during Programming

ΔV_T vs. programming time of the single finger and the multi-finger devices, exhibiting higher programming efficiency achieved in the single finger device, is shown in Figure 8.

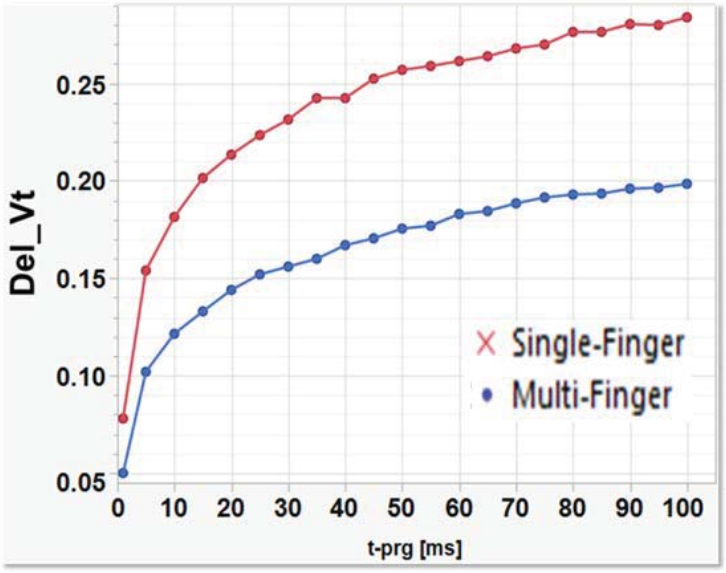


Figure 8: ΔV_T vs. Programming Time for Single Finger and Multi-Finger Devices

As previously stated, the phenomenon of self-heating enhanced charge trapping is equally applicable to planar FET as well as FinFET based devices. To demonstrate the aforementioned, devices fabricated in 14nm bulk FinFET technology were used. Two devices with an equal number of fins (*i.e.*, total effective W_{ch} is identical between the two devices) but different layouts were evaluated. The first device is a '1 gate x 12 fin' device (shown in Figure 9(a)) and the second device is a '2 gate x 6 fin' device (shown in Figure 9(b)). Since the active area-perimeter ratio of the 2x6 device is larger, it has a higher effective thermal resistance. Therefore, lower heat dissipation leads to higher device temperature at high bias conditions during programming. The thermal profile of the two devices during programming is shown in Figure 10. Similarly to planar devices, higher thermal resistance leads to higher programming efficiency in FinFETs. ΔV_T vs. programming time of 1x12 and 2x6 devices, exhibiting higher programming efficiency achieved in 2x6 devices, is shown in Figure 11.

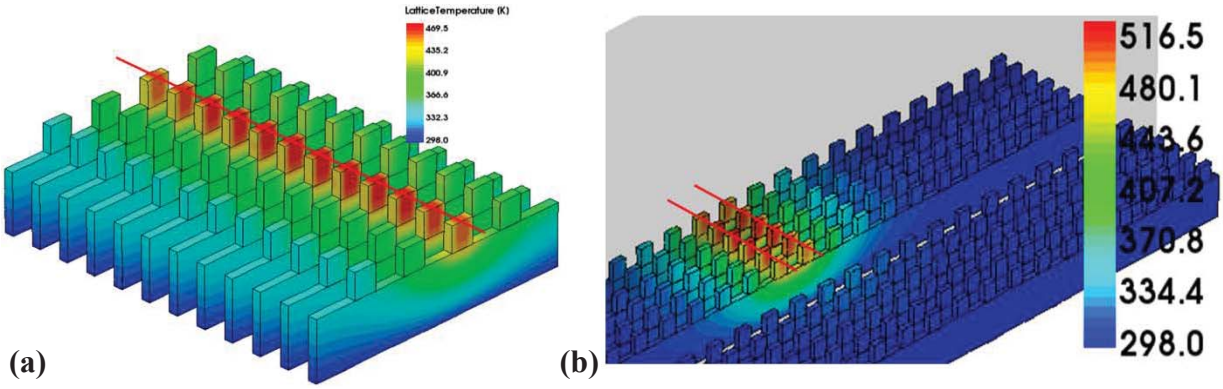


Figure 9: 3D Structure of (a) '1 gate x 12 fin' Device and (b) '2 gate x 6 fin' Device Active Areas

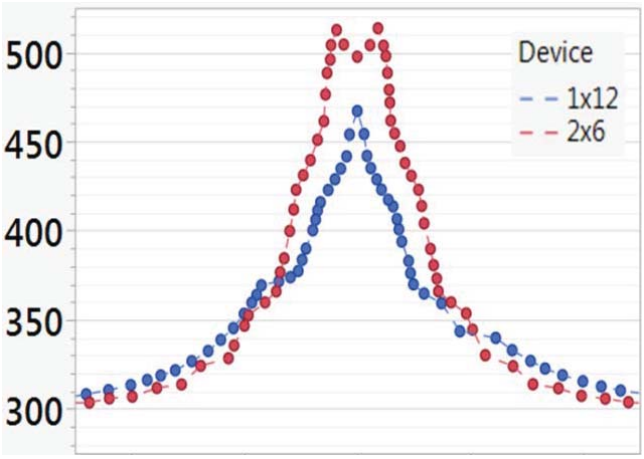


Figure 10: Thermal Profiles of 1x12 and 2x6 Configuration FinFET Devices during the Programming Operation

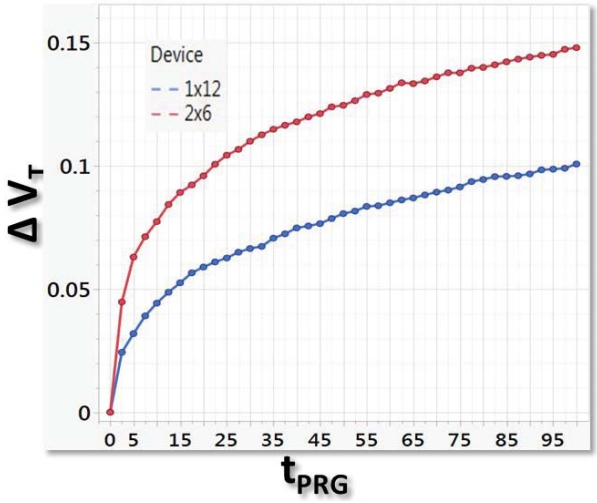


Figure 11: Comparison of ΔV_T vs. Programming Time for the 1x12 and the 2x6 Device

As previously stated, in addition to a significant impact on the programming efficiency of CTTs, thermal resistance also modulates the charge de-trapping behavior, *i.e.*, retention of the trapped charge. To evaluate the charge de-trapping behavior, a set of identical devices is programmed at various fixed V_d to achieve a cumulative ΔV_T of ~225 mV in each device using PVRS. The devices are then stored at an elevated temperature of 85C. Retention of the trapped charge in each of the devices is measured by monitoring the V_T of the device as a function of time. The decrease in V_T (correlated to the loss of trapped charge) is plotted as a percentage of the initial threshold voltage values in Figure 12. It can be observed that retention of trapped charge shows a positive correlation to the stress drain bias V_d.

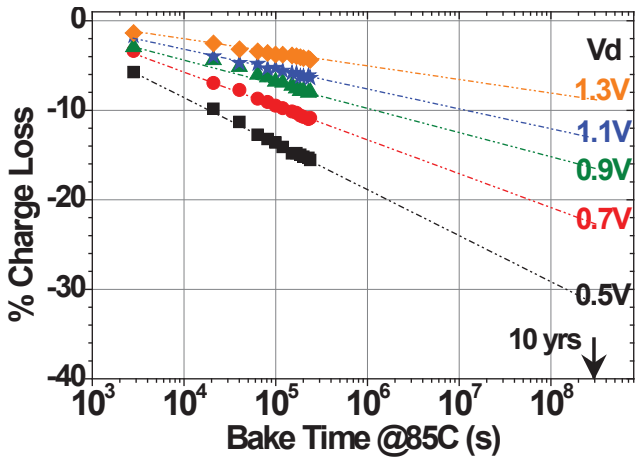


Figure 12: Percentage Charge Loss vs. Bake Time @ 85C, for Identical Devices stressed at various fixed Drain Biases
(W_{ch}=1.2µm, L=20nm)

Another set of devices with different channel width (same length) and different channel length (same width) is programmed using PVRS at V_d=1.5V to achieve a cumulative ΔV_T of ~265mV in each device, and then stored at 85C. The retention of the trapped charge was measured as described above and is shown in Figure 13. As can be seen, trapped charge in wider and shorter devices exhibits higher retention. The enhanced charge retention in wider devices is attributed to higher self-heating. The enhanced charge retention in shorter devices is attributed to a cumulative effect of higher self-heating due to higher power densities as discussed below, and elevated levels of hot carriers due to higher lateral fields.

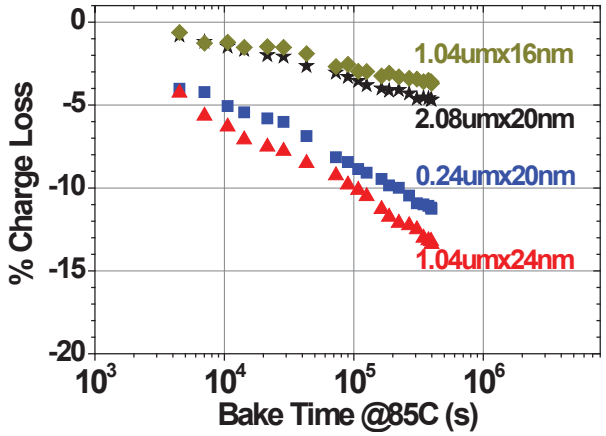


Figure 13: Percentage Charge Loss vs. Bake Time @ 85C for Devices with various Dimensions (W_{ch} x L, as labeled) stressed at V_d=1.5V

Higher stability of charge that is trapped at high temperature (device self-heating induced), as compared to charge trapping at room temperature [1][11], can be attributed to the fundamental nature of charge trapping and de-trapping, which are thermally activated processes. Therefore, capture and emission time of the trapped charge is directly correlated to the activation energy [12], as illustrated in Figure 14. At low temperatures, stable traps with high activation energies (long capture times) require longer time to be filled (shown in Figure 14(a)). Self-heating

induced high temperature enables access to these stable traps in shorter times, and they can be rapidly filled during charge injection (shown in Figure 14(b)). Localization of self-heating leads to rapid cooling (in the nanoseconds range) after the stress conditions are removed, preventing charge de-trapping since activation energy for de-trapping can no longer be achieved, leading to longer emission times and enhanced retention (shown in Figure 14(c)). This understanding is consistent with the known properties of distributed oxide traps such as oxygen vacancies [12].

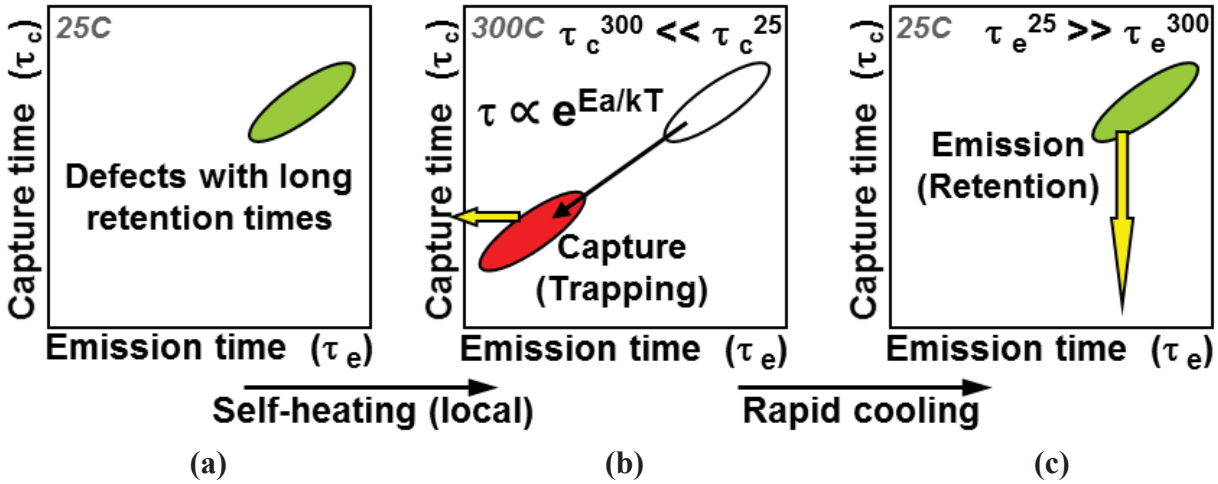


Figure 14: Schematic of 'Capture-Emission Time Maps' for Self-Heating Assisted Charge Trapping (adapted from [12])

(a) Defects with long emission times/good retention also have long capture times, (b) capture times are reduced at elevated temperatures, and (c) rapid quenching retains charge in defects with long emission times at low temperatures.

To understand the charge injection behavior and mechanism(s), charge injection currents during the programming operation are measured at the gate terminal of CTT devices. Measured current for various drain voltage values used during programming, for devices with different R_{th} is shown in Figure 15(a). Similar to the observed trend of ΔV_T, for the same power density, the charge injection current for the single-channel device (higher R_{th}) is considerably higher as compared to the split-channel device (lower R_{th}). However, when plotted as a function of the calculated channel temperature (shown in Figure 15(b)), the charge injection current characteristics of the two devices are very similar. In other words, the magnitude of charge injection current shows a significantly stronger correlation to the self-heating temperature of the device than to the applied power density. This observation reaffirms the conclusion that device self-heating temperature during the programming operation is a significant factor in modulating the charge injection and in turn the charge trapping behavior of CTT devices.

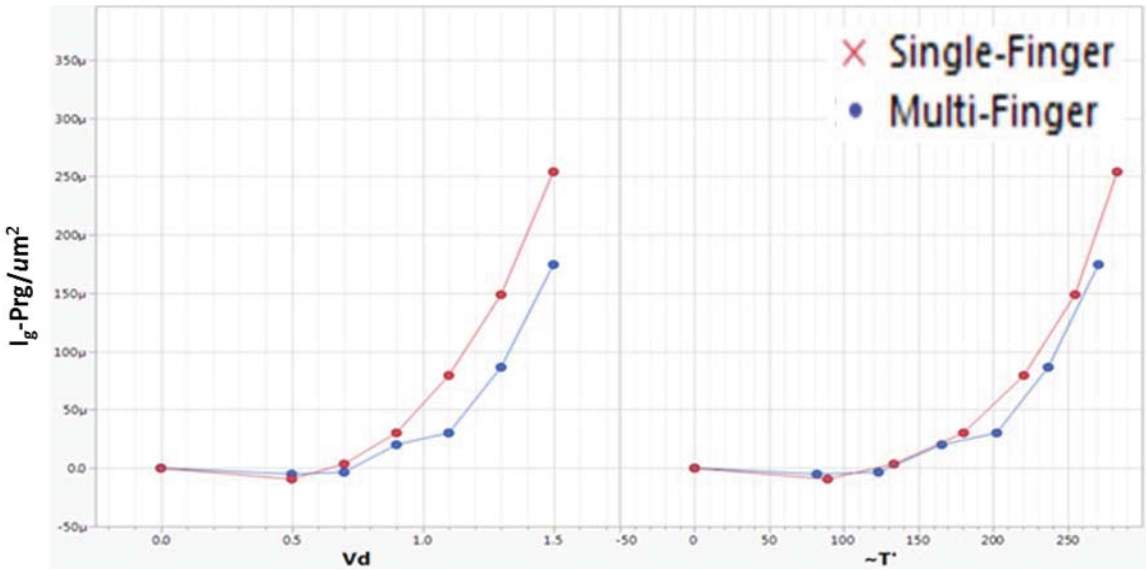


Figure 15: Measured Charge Injection Current during Programming vs. (a) Applied Bias and (b) Channel Temperature (T) during Programming Operation

When the applied gate voltage is lower than the barrier height ϕ_b, the energy offsets between the Si and gate dielectric conduction bands and the electrons tunnel from the Si through a trapezoidal barrier, *i.e.*, direct tunneling occurs. When the applied voltage exceeds ϕ_b, electrons tunnel through a triangular barrier, analogous to field emission leading to Fowler–Nordheim (FN) tunneling. I-V characteristics of direct and FN tunneling are governed by the following relationships

$$ln\left(\frac{I_g}{V_g^2}\right) \propto \begin{cases} \ln\left(\frac{1}{V_g}\right), for\ V_g < V_{trans} \rightarrow Direct\ Tunneling \\ -\left(\frac{1}{V_g}\right), for\ V_g > V_{trans} \rightarrow FN\ Tunneling \end{cases},$$

where, I_g is the gate injection current, V_g is the gate bias during programming, and V_{trans} is the voltage at which the transition from predominantly direct tunneling to FN tunneling occurs.

As evident from the above expressions, a plot of $ln\left(\frac{I_g}{V_g^2}\right)\ vs.\frac{1}{V_g}$ should yield a positive slope in the direct tunneling regime and a negative slope in the FN tunneling regime. The inflection point would be where the charge injection current transitions from being predominantly direct tunneling to predominantly FN tunneling, and Vg=Vtrans at that point. A plot of $ln\left(\frac{I_g}{V_g^2}\right)\ vs.\frac{1}{V_g}$ (shown in Figure 16) for CTT devices programmed at various drain bias voltages (exhibiting different device self-heating temperatures), confirms the aforementioned. Two distinct observations are made: (i) charge injection mechanism appears to be predominantly FN tunneling beyond V_g ~1.6V. Therefore, CTT devices programmed at V_g=2V will have FN tunneling as the predominant charge injection mechanism during the programming operation. (ii)

With increasing V_d and in turn increasing channel temperature due to device self-heating, FN tunneling becomes dominant at relatively higher V_g, *i.e.*, As V_g is increased towards the FN tunneling regime, the direct tunneling remains dominant for longer, which is consistent with the fact that direct tunneling is a much stronger function of temperature as compared to FN tunneling.

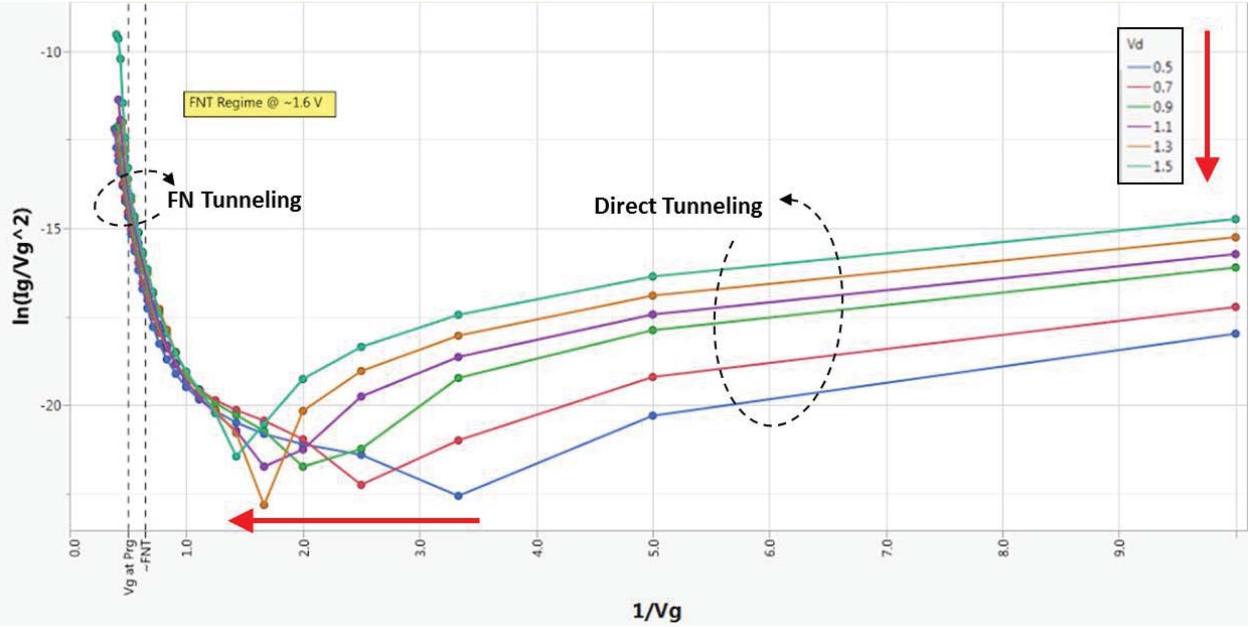


Figure 16: $\ln\left(\frac{I_g}{V_g^2}\right)$ vs. $\frac{1}{V_g}$ for CTTs Programmed at various Drain Bias Voltages

Note that since the gate dielectric that is used is HfO_2 (as discussed earlier, has oxygen vacancies that serve as charge trapping centers), and the channel temperature is considerably high during CTT programming, it is likely that charge injection mechanisms such as Poole-Frenkel (PF) emission and Schottky-Richardson (SR) emission [13] (a thermionic emission of an electron jump over a surface barrier) also exist during the CTT programming operation. Additional experiments and modeling will be required to quantify the relative contribution of each charge injection mechanism.

1.3 Charge Trapping Profile along Device Channel

It was concluded that there is no significant polarity effect (forward vs. reverse mode read) on the mean value of threshold voltage and saturation current of CTT devices (shown in Figure 17 and Figure 18). These results suggest that, while some asymmetry might be present, the trapped charge is fairly uniformly distributed along the channel. The corresponding stochastic variation results in 2.8% standard deviation for normalized deltas between forward and reverse polarity (shown in Figure 18).

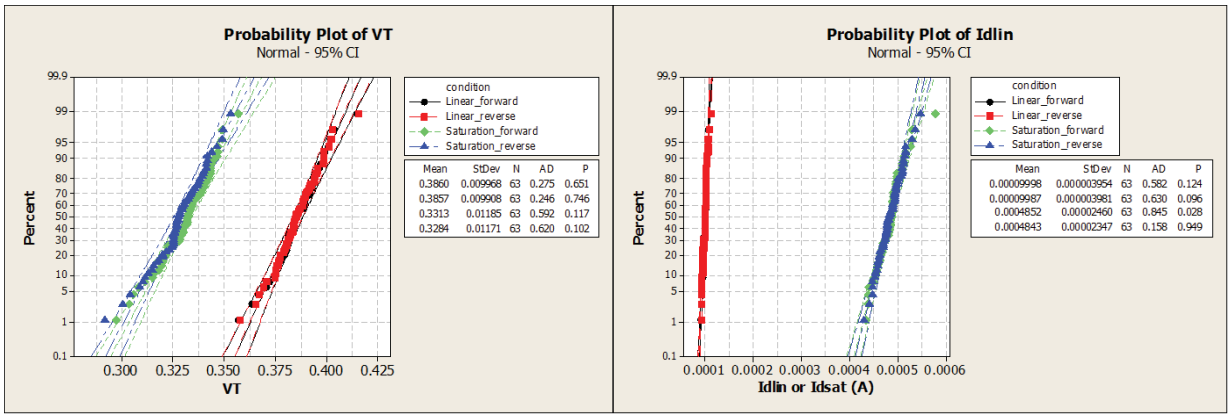


Figure 17: Reverse vs. Forward Mode Distribution for (a) V_T and (b) Linear and Saturation Mode Ion

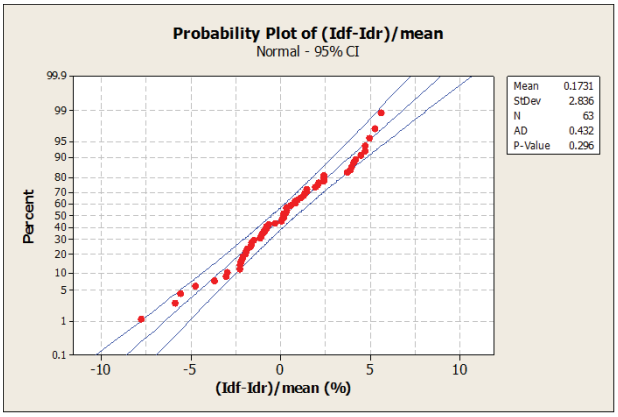


Figure 18: Stochastic Variation in Standard Deviation for Normalized Deltas between Forward and Reverse Mode Reads

Post-programming forward vs. reverse mode transconductance (Gm) measurements do not show any significant asymmetry, as shown in Figure 19, further indicating that the trapped charge distribution is fairly uniform along the channel.

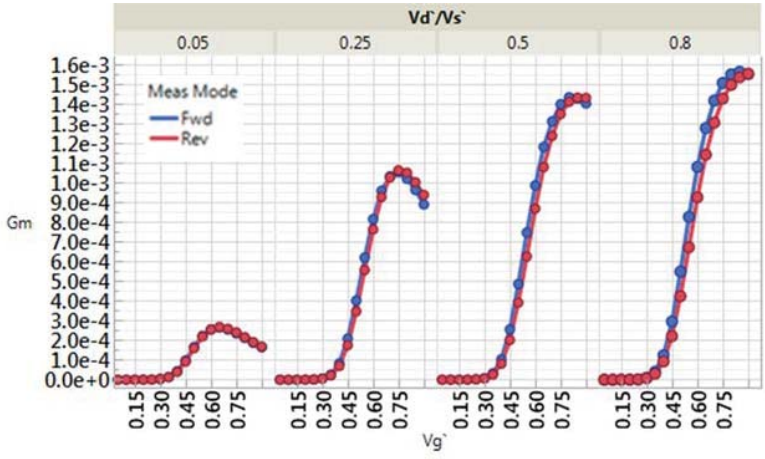


Figure 19: Device Transconductance vs. V_g (at various V_d/V_s values) for Forward and Reverse Mode Reads Overlaid

1.4 Verification that Trapped Charge is not exclusively on one Side of the Channel (or the sidewall) via TCAD Simulations

A dual gate 20nm metal-oxide-semiconductor field-effect transistor (MOSFET) was simulated to study the effect of asymmetric charge trapping in the gate dielectric, as shown in Figure 20.

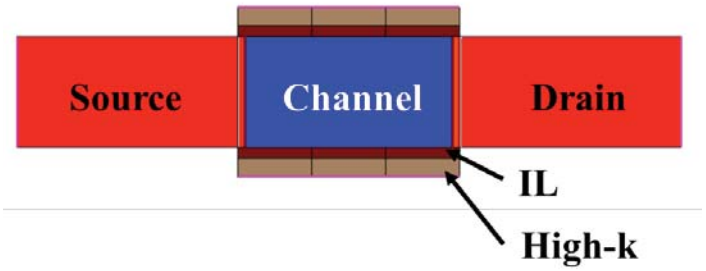


Figure 20: Top Down View of Device used for Simulation of Asymmetric Charge Distribution in the Gate Dielectric

The trapped charge was placed only on 1/3rd of the channel with its position skewed towards the source side, as shown in Figure 21.

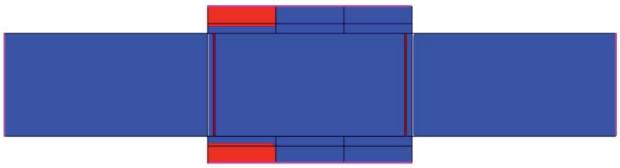


Figure 21: Top Down View of Device Symmetric Charge Distribution in the Gate Dielectric
(Charge shown in red)

I-V characteristics (shown in Figure 22) of the evaluated device show that for this trapped charge profile, there is *significant* asymmetry between forward and reverse mode read. A ΔV_T of 100mV in forward mode read and ~35mV in reverse mode is observed. In other words, forward mode ΔV_T is 300% higher than that in reverse mode, which is an expected result. Our experimental data shows a maximum of approximately 20-25% ΔV_T asymmetry, further suggesting that the trapped charge is fairly uniformly distributed within the gate dielectric along the channel of the CTT.

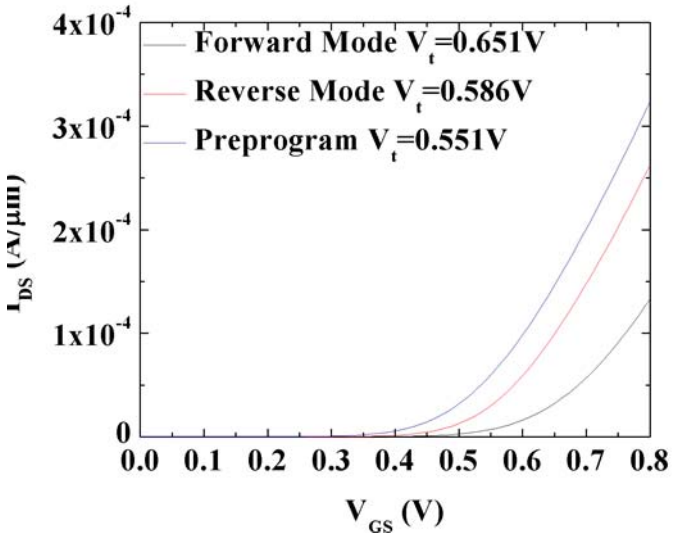


Figure 22: Forward vs. Reverse Mode I_d-V_d Sweeps for Asymmetric Charge
(Placed above 1/3rd of the channel)

1.5 Application of CTT Devices as Multiple-Time Programmable Memory Elements

To understand the fundamental physical mechanisms and study the program/erase (P/E) characteristics of CTTs for their application as MTPM elements, P/E cycling of the devices was performed using the PVRS technique. 10ms V_g pulses of increasing magnitudes in increments of 10mV were applied, as demonstrated in Figure 23, for various fixed programming V_d values. The very first program operation, referred to as 'Initialization', is unique. This is followed by an erase ('ERS') operation using negative PVRS and then a re-program ('PRG') operation. The source and drain are typically grounded during erase operations. The observed behavior reveals the presence of three distinct V_d-dependencies which can be exploited in a CTT for an MTPM application. (i) During 'initialization' ΔV_T exhibits a strong dependence on V_d. For higher V_d, equivalent ΔV_T values are achieved at significantly lower V_g. This effect is due to a combination of enhanced trapping and trap creation in the HfO_2 at higher V_d (stronger device self-heating). (ii) For devices programmed at higher V_d, longer time and/or larger negative Vg values are needed to de-trap the charge. Charge trapping at high temperature (stronger self-heating at high V_d) is more stable and it is more difficult to erase the devices. The slight ΔV_T difference between the end of the 'Initialization' cycle and beginning of the 'ERS' cycle is believed to be caused by fast de-trapping of the small fraction of unstable trapped charge in each case, followed by no further de-trapping until a certain negative bias is applied during the 'ERS' cycle. The magnitude of this small ΔV_T difference is inversely proportional to the programming V_d, which is again consistent with the relation between programming V_d and overall trapped charged stability. (iii) The charge trapping behavior changes after the 'Initialization' operation due to the creation of new traps [4][5] allowing for subsequent programming ('PRG') to the same ΔV_T at lower V_g. This phenomenon has also been reported in [14], where an increased rate of charge trapping for pre-stressed devices is attributed to new trap creation during the charge injection process. To verify the above and compare self-heating enhanced charge trapping to conventional bias temperature-instability, PVRS sweeps were also performed with V_d=0V (shown in Figure 23 inset). Without the effect of self-heating, (i) for the same Vg values, the obtained ΔV_T is

relatively very small, (ii) the ΔV_T is fully recoverable (*i.e.*, traps discharge easily), and most importantly (iii) the charge trapping behavior does not change subsequent to the first cycle and is repeatable for many cycles, indicating that creation of additional traps is minimal.

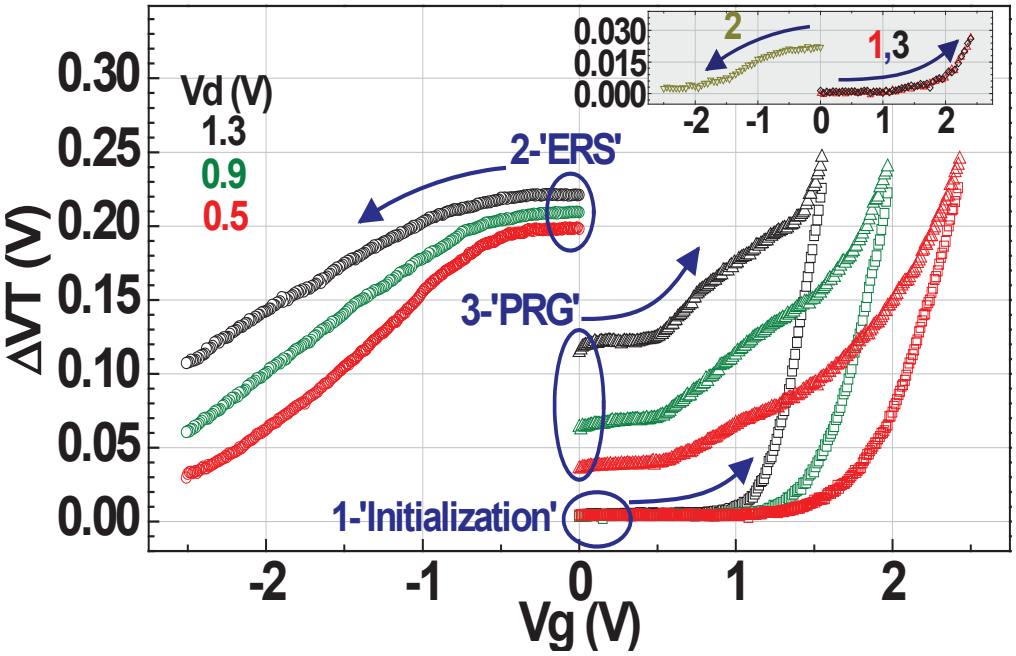


Figure 23: Measured ΔV_T during 1-'Initialization', 2-'ERS', and 3-'PRG' Cycles for various V_d Values using PVRS

Inset shows ΔV_T for V_d =0V PVRS stress (bias temperature-instability).

There is a tradeoff between trapped charge retention, ΔV_T window, and erase time/voltage needed. Higher programming V_d leads to more stable V_T shifts (better retention) as demonstrated and discussed earlier, but it will take longer time and/or higher voltage to erase the cells. In other words, for a given erase time/voltage constraint, the ΔV_T window will be smaller if higher programming V_d is used. Thus, it is important to optimize the operating conditions of the memory cells. Typically, longer erase time, as compared to programming time, is needed to avoid under-erasing, and shorter programming time, as compared to the time in the 'Initialization' operation, is needed to avoid over-programming, to achieve a sufficiently large 'memory window'. It is also advantageous to perform 'Initialization' at a higher V_d than that subsequently used during the 'PRG' operations to avoid over-programming in subsequent P/E cycles. At the same time, the selected V_d for programming must be high enough for the trapped charge to have acceptable retention for the memory application. While this discussion provides a general guideline, detailed optimization will depend on device geometry, layout, and gate stack properties, all of which affect the charge trapping behavior.

To demonstrate the importance of optimizing P/E conditions for the memory application, CTT devices were cycled 20x using un-optimized P/E conditions (*i.e.*, P/E conditions were not optimized to avoid over-programming and under-erasing) and compared to devices with optimized P/E conditions ($V_{d\text{-}PRG}$ slightly lower than $V_{d\text{-}INIT}$, limited number of 'PRG' pulses to avoid over-programming, and longer erase time during the 'ERS' operation to achieve maximum ΔV_T recovery). Post-program and post-erase ΔV_T values for the devices in each case are shown

in Figure 24. By optimizing P/E conditions, over-programming and under-erasing with P/E cycling (which causes the 'memory window' to dynamically drift higher, resulting in a shrinking read-margin for the "erased" state with respect to a fixed reference read voltage, as seen with un-optimized P/E conditions) can be avoided, resulting in significant improvement in the endurance of the memory cells. In functional memory arrays, program and erase 'verify' schemes are used to further optimize the P/E operations. Post-program and post-erase ΔV_T values for devices that were cycled 800x using optimized P/E conditions are shown in Figure 25. It can be seen that even after 800 cycles a stable memory window exists.

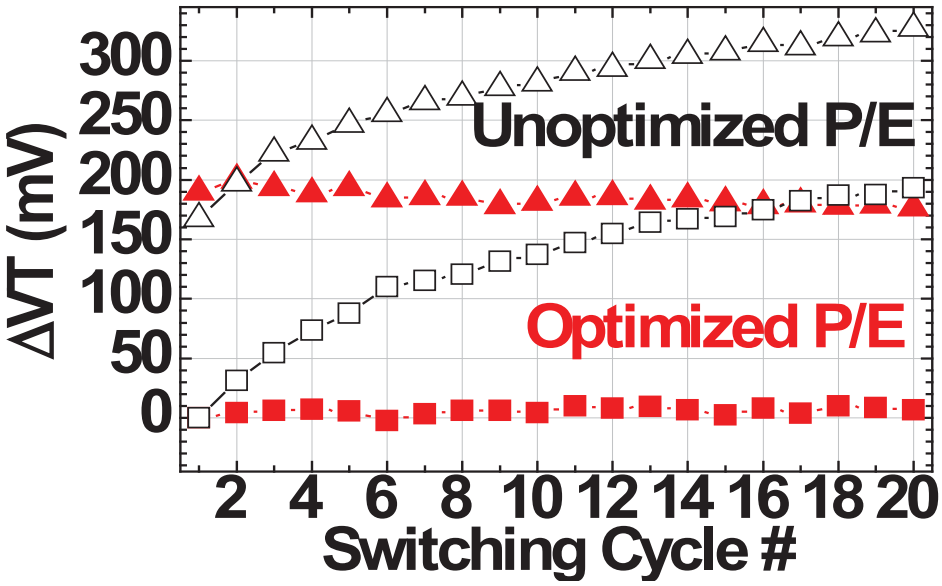


Figure 24: Memory Window vs. Switching Cycle Number Comparison between Un-optimized and Optimized P/E Conditions

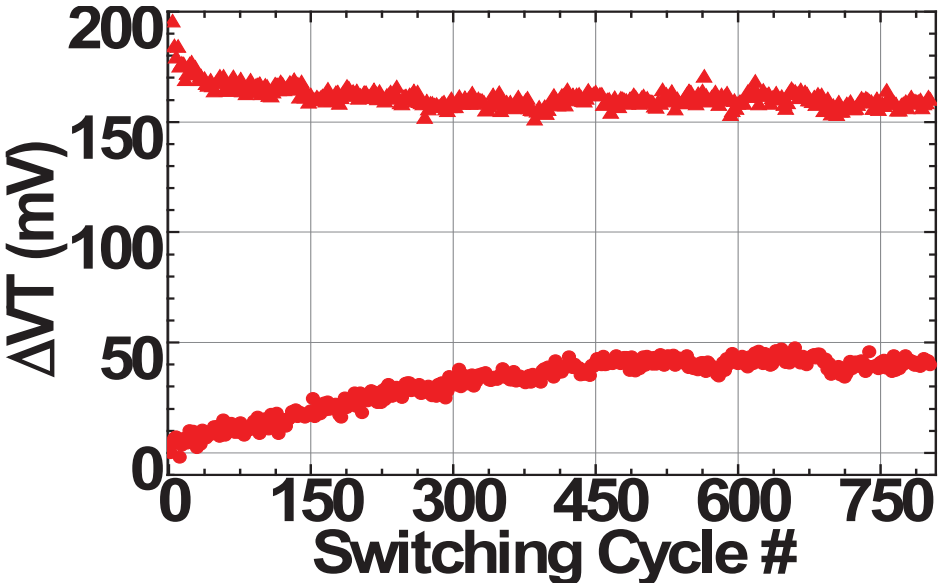


Figure 25: P/E Cycling of CTTs using Optimized Operation Conditions

Fully functional CTT memory arrays have been successfully integrated in 22nm SOI planar, 14nm SOI FinFET, and 14nm Bulk FinFET technology platforms. Efforts to demonstrate feasibility of CTT based eNVM in 7nm bulk and 12nm fully depleted silicon on insulator (FDSOI) technologies are ongoing.

Bitmaps of fully-functional CTT memory arrays fabricated in 22nm SOI planar, 14nm SOI FinFET, and 14nm bulk FinFET production technologies are shown in Figure 26. Details about circuit design and sensing techniques implemented in this technology have been discussed in [15].

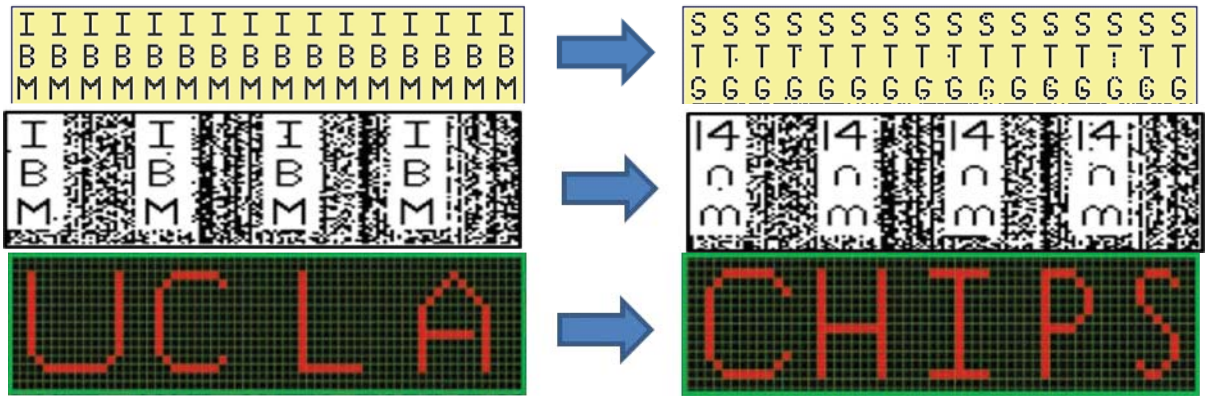


Figure 26: Bitmaps of a Fully Functional CTT Memory Arrays integrated in (top) 22nm SOI Planar, (middle) 14nm SOI FinFET, and (bottom) 14nm Bulk FinFET Production Technologies

The bit patterns on the left are written followed by an erase and re-write of an alternate bit pattern.

1.6 Bit-cell Architecture

Two approaches are possible for a CTT bit-cell: single-cell and twin-cell. In the single-cell bit approach, a universal reference bit is used for all bit-cells in a bitline, while in the twin-cell approach, each bit-cell includes a dedicated reference bit. While the single-cell approach is clearly favorable for achieving higher array densities, the twin-cell approach can be utilized to improve sense margin and make the circuit more immune to device-to-device variability which is inherent to any technology platform. A schematic of a twin-cell is shown in Figure 27. Details on circuit design and operation have been discussed in [15][16].

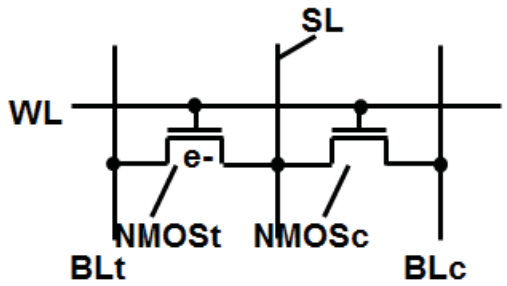


Figure 27: Twin-Cell Architecture of a CTT Bit-Cell

1.7 Reliability Considerations

Gate leakage of the CTT device increases with increasing ΔV_T. This is not unexpected since an increase in the trap density within the high-k layer leads to an increase in trap-assisted tunneling (TAT). Since the highest vertical field across the gate dielectric during PRG is at the source side (V_g=high, V_s=0V, V_d=high), as expected, post-PRG gate leakage is higher in the source side, as compared to the drain side. We have confirmed the preceding with reverse vs. forward mode reads. A high V_{gs} (when V_{gd}=0V) results in a higher gate leakage as compared to the same V_{gd} (when V_{ds}=0V), as shown in Figure 28.

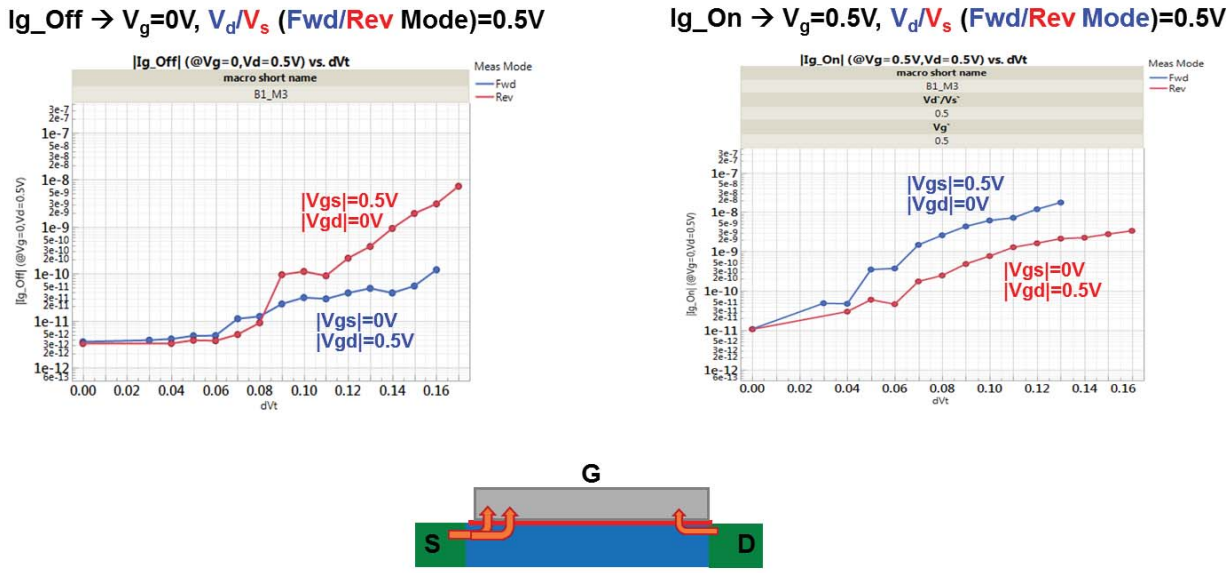


Figure 28: Gate Leakage Current vs. ΔV_T observed during Device Off-State (left) and On-State (right)

The schematic shows the overall message that the source side is leakier after PRG, as it experiences the highest vertical field during the PRG operation.

As can be seen from the Ig_off and Ig_on data (shown in Figure 29), Ig_off is higher in the reverse mode and Ig_on is higher in the forward mode. Since the sum of Ig_off of all the devices that share the same bitline impacts the total signal-to-noise-ratio in the array, while Ig_on impacts only a single device, read conditions favoring lower Ig_off are generally preferred. Additionally, Ig_on at any read condition is several orders of magnitude smaller that the channel current. However, the optimized read conditions will depend on the particular array design, sense amplifier design, and application. Work to model and develop a methodology for optimization of read conditions is ongoing.

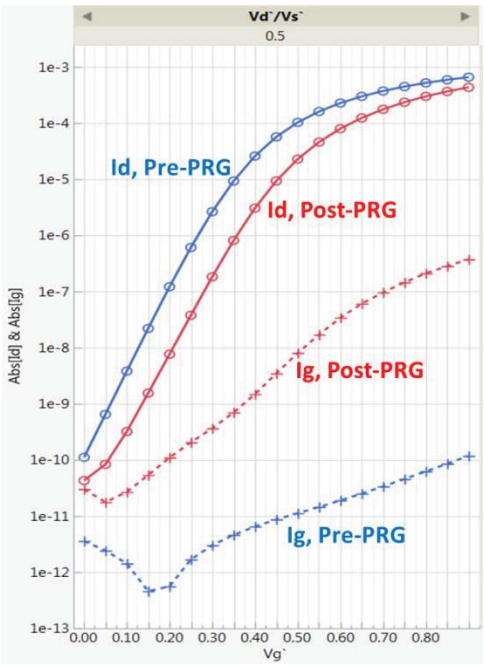


Figure 29: I_d-V_g and I_g-V_g Sweeps of Pre-PRG and Post-PRG Devices
Ig_on at any read condition is several orders of magnitude smaller that the channel current.

We observed a small reduction in the device Gm post programming, indicating that some of the trapped charge may be at or near the channel-gate dielectric interface causing coulomb scattering of carriers. The Gm reduction in 14nm Bulk FinFETs (~5%) is slightly larger than what was observed in 22 SOI (~3%). However, a relatively small reduction in Gm is not expected to have any consequential effect on the operation of CTTs as digital memory devices. Gm vs. V_g (at a read V_d of 0.5V) for pre-PRG and post-PRG devices (ΔV_T ~100-120 mV) are shown in Figure 30.

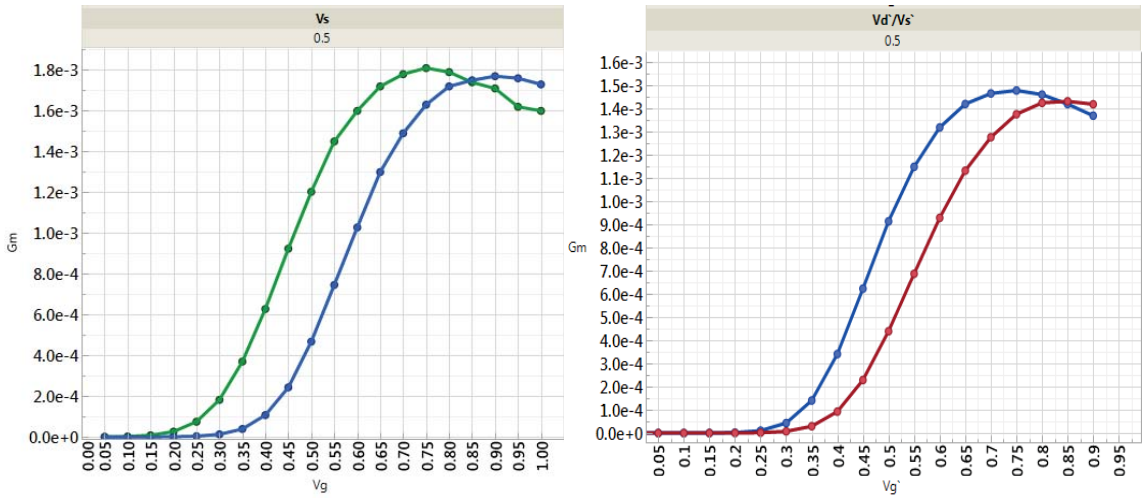


Figure 30: Gm vs. V_g (at a read V_d of 0.5V) for Pre-PRG and Post-PRG Devices (ΔV_T ~100-120 mV)
Left panel shows data from 22nm SOI CTTs and right panel shows data from 14nm bulk FinFET CTTs.

1.8 Comparison between CTTs in various Technologies

As discussed and demonstrated earlier, CTT based eNVM has been successfully integrated in various technology nodes including 32nm partially depleted silicon on insulator (PDSOI) planar, 22nm PDSOI planar, 14nm SOI FinFET, and 14nm bulk FinFET. Development of CTT based eNVM technology in 7nm bulk FinFET and 12nm FDSOI planar technology platforms, is ongoing.

We have found that the fundamental device physics and device operation are similar across all technology platforms. However, there are a few differences between technology nodes that are of interest and are being investigated:

(i) Programming efficiency: we observe that 14nm bulk FinFETs program relatively slower as compared to 22nm SOI CTTs. While several reasons could be possible for this difference, SOI vs. bulk is one of the most obvious one. Intrinsic self-heating enhanced charge trapping memory devices, *i.e.*, CTTs are implementable in SOI as well as bulk FinFET technologies as self-heating in bulk FinFETs, while generally less than SOI FinFETs, is comparable to SOI planar devices and increases considerably with scaling [17][18]. However, SOI devices tend to have a relatively higher self-heating as compared to bulk devices, which is a possible reason to the reduced efficiency of charge trapping.

Gate dielectric thickness is another parameter that is possibly responsible for the different programming efficiency. Due to several architectural differences between planar SOI and bulk FinFET technologies, decoupling the impact of a gate dielectric thickness from other architectural and parametric differences is difficult. However, the gate dielectric in the technology platform used for 14nm bulk FinFET CTTs is slightly thicker than that of the technology platform used for 22nm PDSOI CTTs, which is a possible reason for the relatively lower programming efficiency in 14nm bulk FinFET CTTs. Both 14nm FinFET and 22nm PDSOI CTTs have similar device dimensions, therefore device dimensions can be ruled out as having an impact on the difference in programming efficiency.

To verify the impact of gate dielectric thickness on the programming efficiency of CTTs, the effect of Post Deposition Anneal (PDA) on the programming efficiency of CTTs in the 14nm bulk FinFET technology is studied. It is found that indeed PDA temperature significantly modulates One Time Programmable Memory (OTPM) program efficiency. Lower PDA exhibits higher drive currents and programming efficiency. Higher drive currents and programming efficiency with lower PDA temperature consistent with thinner InterLayer (IL), allowing more charge to be injected in the high-k dielectric during programming. Shown in Figure 31 is the programming efficiency (ΔV_T for same programming conditions) as a function of the PDA temperature.

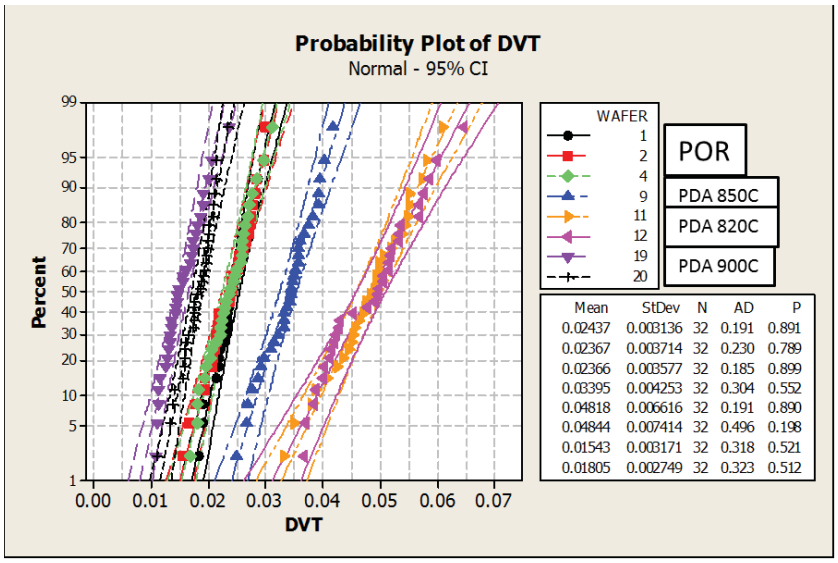


Figure 31: Impact of PDA Temperature on Programming Efficiency in 14nm Bulk FinFET CTTs

(ii) Gate leakage vs. ΔV_T: an increase in gate leakage with increasing ΔV_T is observed in CTTs across all technology nodes, which is not unexpected as an increase in the trap density in the high-k layer leads to an increase in TAT. In addition, and as discussed and demonstrated earlier, as long as the signal-to-noise ratio is large enough for all bits to be read without error, gate leakage should not impact the functionality of the digital memory arrays. As also discussed earlier, read techniques are available to minimize the impact of gate leakage. Further understanding and optimization of operation conditions to minimize gate leakage is ongoing. However, the increase in gate leakage in 14nm FinFET CTTs is observed to be somewhat higher as compared to 22nm and 32nm planar CTTs. A comparison between Ig_on vs. ΔV_T for 22nm PDSOI and 14nm bulk FinFET CTTs is shown in Figure 32.

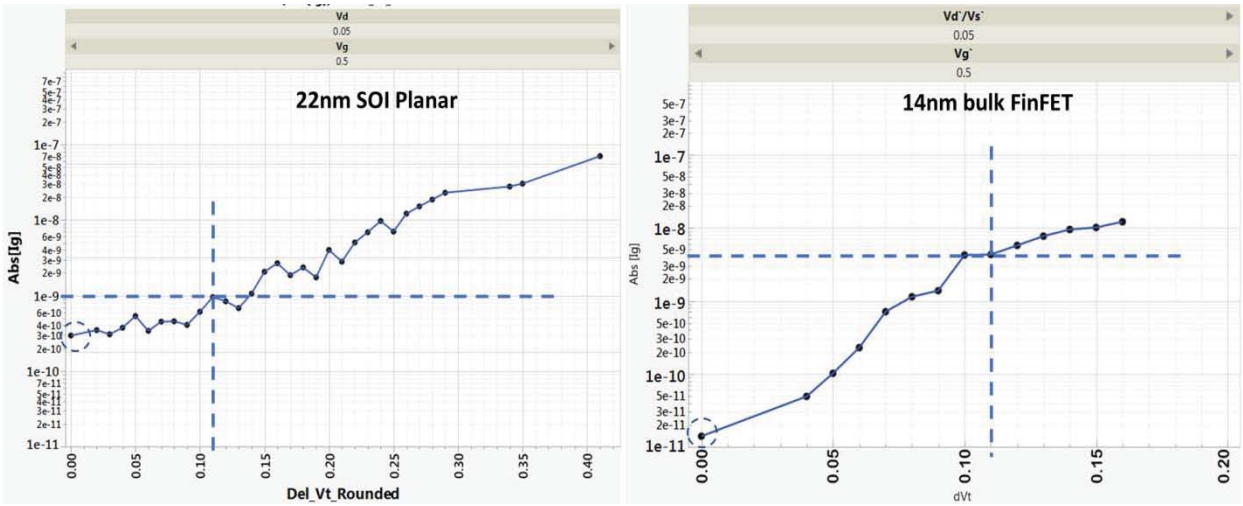


Figure 32: Comparison of Gate Leakage Current vs. dVt created for 22SOI Planar (left) and 14nm bulk FinFET (right) Devices

While further work to fully understand the differences is ongoing, there are several reasons that might be the cause for this difference in gate leakage:

(a) One significant difference between the 32nm PDSOI and 22nm FDSOI vs. 14nm FinFET processes, is gate-first vs. gate last processes, respectively. The gate dielectric in gate-first process experiences the entire thermal budget while the gate dielectric in gate-last process experiences only the post deposition anneal. The resultant difference in the quality of the gate dielectric is possibly one reason for the different gate leakage behaviors. Recently, we have learned that 22nm FDSOI, which is also a gate-first process, shows lower post-programming gate leakage similar to 22nm SOI, which further indicates that the gate-first vs. gate-last process might be a driving factor in the post-programming gate leakage.

(b) Effects due to architectural differences such as gate edge fringing fields could possibly cause slightly higher gate leakage, *i.e.*, the same amount of traps can result in a higher leakage current due to higher field at the replacement gate edges.

1.9 Summary and Conclusions

Charge trapping behavior in HKMG SOI CMOS devices has been analyzed and it is found that not only the magnitude, but also the stability of the trapped charge increases with device self-heating during the charge injection process. The same magnitude of charge trapping can be achieved in much shorter times and has higher stability when the devices are stressed at higher drain bias or higher self-heating conditions. Device geometry and layout significantly modulate R_{th} and in turn the self-heating and charge trapping behavior.

The promising charge retention characteristics that have been demonstrated (extrapolated accelerated charge loss of <10% after 10 years at 85C) make it feasible to utilize HKMG CMOS devices as memory elements for scalable fully integrated system-on-chip applications such as embedded non-volatile memory. These devices can also serve as a potential alternative to existing one-time programmable technologies like eFUSE [19] for yield improvement, performance tailoring, field configurability, and data security enhancements such as on-chip reconfigurable encryption key storage, firmware storage, and Chip IDs.

We reiterate that, as demonstrated and discussed here, CTT based eNVM memory can be realized in bulk as well as SOI technologies without any additional processing or masks. The CTT device operates at voltage and power levels that are logic compatible (~2V, ~4mW peak power), is multiple-time programmable, offers an SRAM type high-density, and scalable.

Integration of functional memory arrays has been realized on 32nm PDSOI planar, 22nm PDSOI planar, 14nm SOI FinFET, and 14nm FinFET production technology platforms, and effort to demonstrate feasibility of CTT based eNVM in 7nm bulk and 12nm FDSOI technologies is ongoing, demonstrating the robustness and commercial potential of this technology.

2. CTT for Analog Memory

We have so far discussed CTT characteristics mainly for digital memory applications. In the next sections of this report, we will describe how CTTs can be used for analog memory – the synapses in particular − in a neuromorphic system.

2.1 Use of CTT as an Analog Memory

2.1.1 Channel Conductance at a given Bias as the Synaptic Weight

A very important computation in any neural network is to calculate the weighted sum of the inputs, where the inputs are weighted by the synaptic strength. In a hardware implementation, the input conveniently takes the form of a voltage. If the conductance of a certain device represents the synaptic weight, then the current will automatically become the weighted input. In the context of CTT devices, the channel conductance of a CTT, at certain bias voltages V_G and V_D, can be used as the synaptic weight. This concept is schematically illustrated in Figure 33(a). In the long-term depression (LTD) regime, for example, short trapping pulses (~20 μs) are applied to the gate with a drain bias, increasing the V_T of the CTT (shown in Figure 33(b)). The channel conductance at a given bias, *e.g.*, $V_G = 0$ and $V_D = 50$ mV, is decreased. On the other hand, in the long-term potentiation (LTP) regime, de-trapping pulses are applied to the CTT and the channel conductance increases (shown in Figure 33(c)). Considering that very short pulses can induce very fine V_T change and therefore very fine conductance change, CTT devices have the potential to be used as analog synapses.

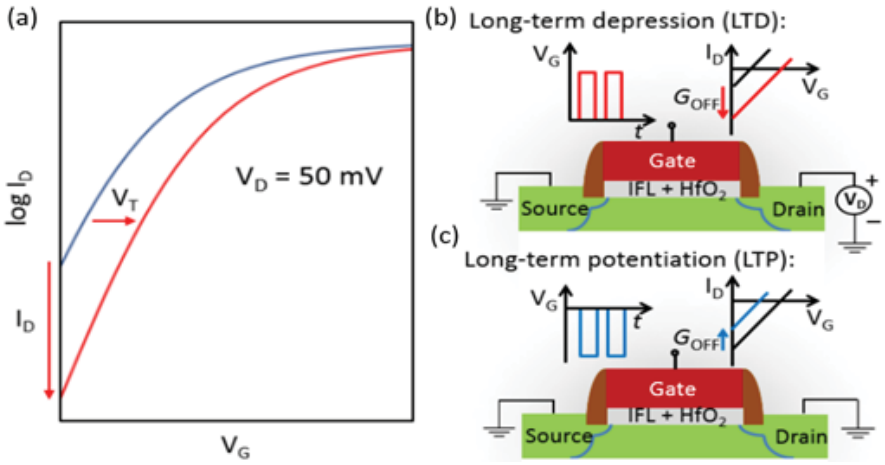


Figure 33: (a) Use of Channel Conductance a certain Bias (V_G and V_D) as the Synaptic Weight, (b) Trapping Pulses increase V_T and decrease the Conductance, and (c) Detrapping Pulses decrease V_T and increase the Conductance

2.1.2 Different from Memristors: Extra Knobs of V_G and V_D

One of benefits of using a transistor (such as a memristor or a phase-change memory) rather than a resistor is that the channel conductance is not determined exclusively by V_T of the device, but also by V_G and V_D bias voltages. This enables a wide range of available weights, providing flexibility in the circuit and system design.

2.1.3 Spike-Timing Dependent Plasticity

The CTT can be employed to realize plastic synapses that possess spike-timing dependent plasticity (STDP) [20]. STDP is a key memory and learning mechanism in biological synapses: with respect to Figure 34(a), if the pre-synaptic neuron repeatedly fires right before the post-synaptic neuron does, the connectivity between the two neurons is strengthened (LTP); however, if the pre-synaptic neuron repeatedly fires right after the post-synaptic neuron does, the connectivity is weakened (long-term depression, LTD). During the update of the weight, when the drain is biased, and the source and gate are subjected to appropriate pre- and post-synaptic neuron pulses (shown in Figure 34(b)), the change in the channel conductance is modulated by the timing difference ($\Delta t = t_{pre} - t_{post}$) between the two pulses, mimicking the STDP behavior. A detailed breakdown for the two cases when $\Delta t > 0$ and $\Delta t < 0$ is shown in Figure 35. In both cases, time segment 3 is the main programming pulse. During experiments, all devices were initially programmed to have a moderate V_T shift by applying a 2.3 V gate pulse for 0.6 ms with a drain bias of 1.5 V before applying the pre- and post-synaptic pulses. Plotted in Figure 36 is the measured conductance change as a function of Δt after synaptic pulses with parameters of V_1 = 1.3 V, V_p = 2.6 V, t_1 = 1ms, and S = 1.3V/13ms (similar to the ones used in Figure 34(a)) were applied to the device.

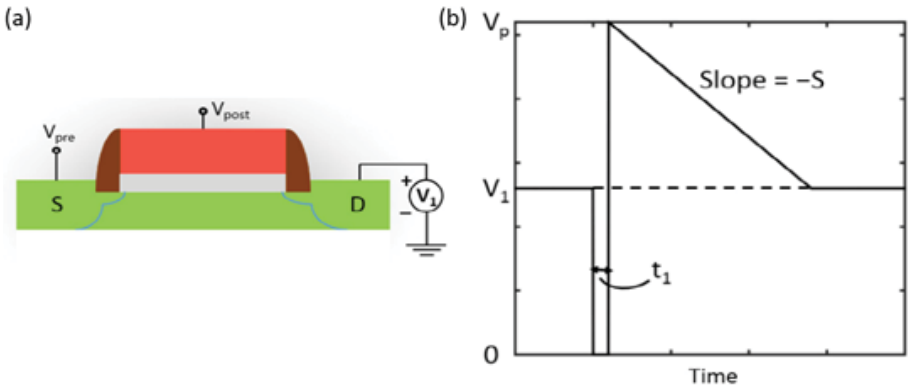


Figure 34: (a) Implementation of CTT as a Plastic Synapse and (b) Example Voltage that is applied as the Pre- and Post-Synaptic Pulses

For (a) the drain is biased at V1, and the source and gate are connected to the pre- and postsynaptic neuron axons, respectively

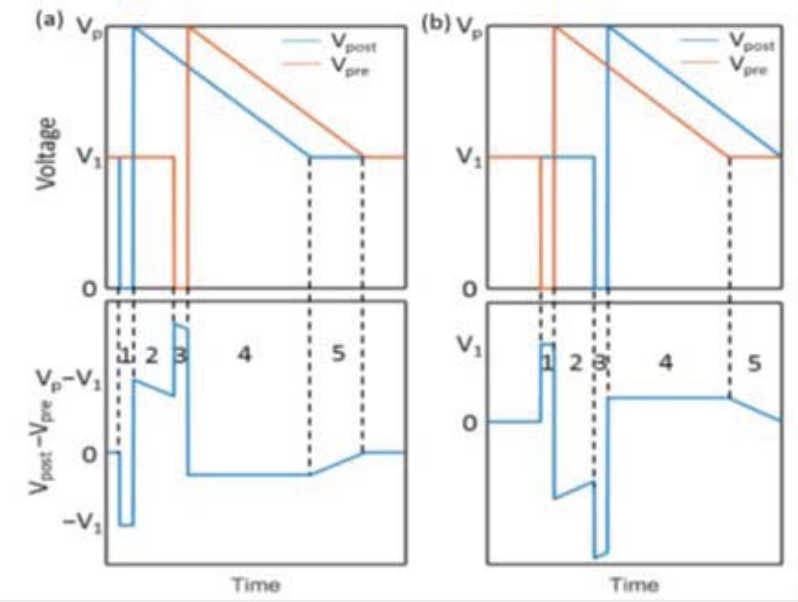


Figure 35: Pre- and Post-Synaptic Neuron Pulses applied to the CTT when (a) tpre−tpost > 0 and (b) tpre−tpost < 0

Shown on the bottom of each plot is the difference between the two pulses. The time segment 3 is the main programming pulse in both cases.

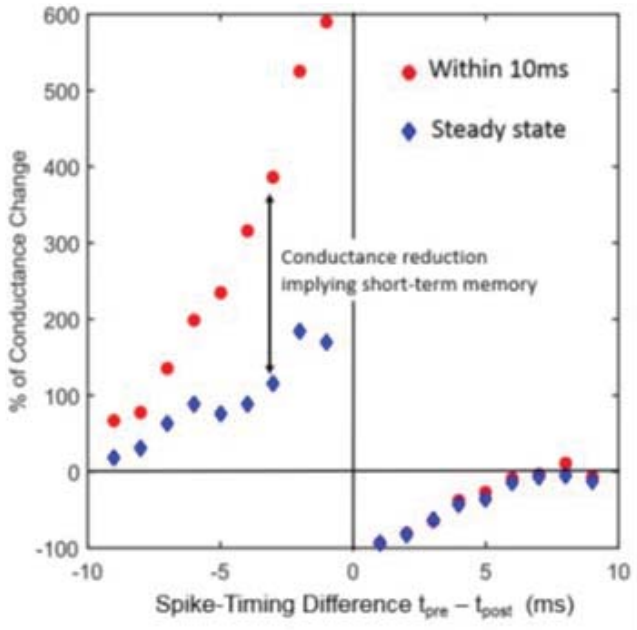


Figure 36: STDP behavior Demonstrated in a 22nm SOI CTT

The conductance of the synapse can be tuned over a range of more than 100x in increments of sub-5mV. For a maximum ΔV_T of 200 mV, at least 100 very fine steps in conductance can be obtained. Note that in the LTP regime, the conductance is initially high immediately after the pulses are applied and then decays to a stable value, indicating the CTT can also be used as a short-term memory device.

2.2 Fine-Granularity Memory using CTT

2.2.1 Experimental Setup

N-type CTTs with an interfacial layer (IFL) SiO_2 followed by an HfSiOx layer as the gate dielectric are used in this study. It should be noted that, although this demonstration features planar SOI devices, the mechanisms apply to bulk substrates/FinFETs as well. The subthreshold OFF-conductance (G_{OFF}) of the CTT at V_{DS} = 50 mV and V_{GS} = 0 is used as the synaptic weight throughout this Report. In the operation of a CTT-based synapse, its G_{OFF} is modified by changing the amount of charge trapped in the high-k layer and thus shifting the V_T of the transistor. In the LTD regime, a positive gate pulse is applied and electrons are trapped into $HfSiO_x$ through the IFL, increasing V_T and decreasing G_{OFF} (Figure 37(a)); in the LTP regime, a negative gate pulse is applied and trapped electrons tunnel back into the channel, decreasing V_T and increasing G_{OFF}, as shown in Figure 37(b). In our experiments, a CTT is first pre-programmed to an intermediate starting state by applying a gate pulse of 2.5 V for 60 µs with V_D = 1.3 V. The device subsequently goes through four cycles: two LTD and two LTP cycles, with 256 trapping or de-trapping pulses in each cycle. In the LTD cycle, G_{OFF} is decreased by a 20-µs, 2.5 V gate pulse with V_D = 1.3 V; in the LTP cycle, G_{OFF} is increased by a 50-µs, −2.6 V gate pulse with zero drain bias.

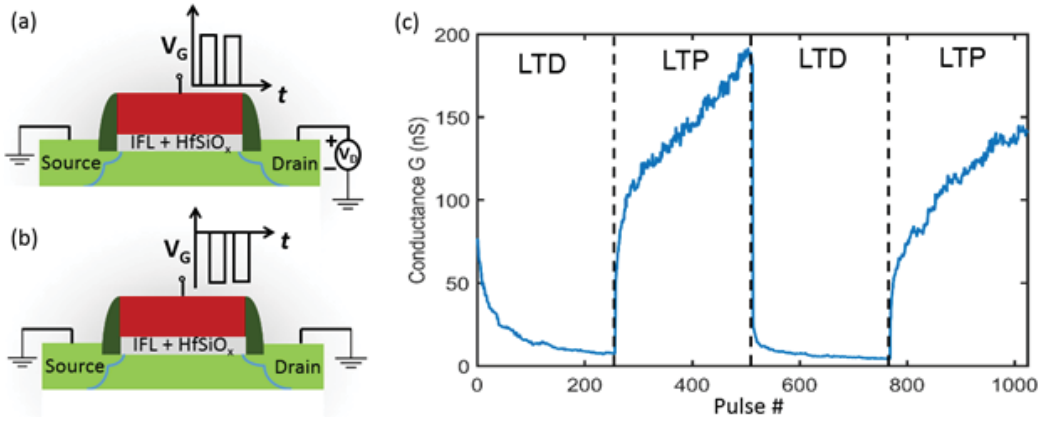


Figure 37: Configurations of the CTT in (a) LTD and (b) LTP Regimes; (c) Reversible and Reproducible Device Conductance change through Four Cycles

2.2.2 Results

The resulting G_{OFF} is shown in Figure 37(c) where a reversible and reproducible modification of synaptic weights can be observed. Over 200 levels are achieved with a very fine resolution of less than 1 nS.

The energy consumption in the LTP regime is minimal since it is only due to electrons being detrapped from the high-k layer. In the LTD regime, the energy dissipation is mainly through the channel current because of the drain bias; it is given by $E = V_{DS} \int I_D dt$ where I_D is the channel current. For a device with a W/L = 20 nm / 20 nm and programming conditions given above, E is estimated to be 0.5 nJ. This is a reasonable value compared to the range of pJ to hundreds of nJ reported for many other synapse candidates [21].

2.2.3 Weight-Dependent Plasticity

An important characteristic of CTTs when used as analog synapses is the weight-dependent plasticity: at different G_{OFF}, the effect of programming pulses on G_{OFF} is different. The weight-dependent plasticity is also found in biological synapses, and might be interesting to emulate the brain. Shown in Figure 38(a) is the relative G_{OFF} change as a function of G_{OFF} itself when five trapping and de-trapping pulses as specified above are applied. It is observed that, in the LTP regime, the relative G_{OFF} increase is smaller when the initial G_{OFF} is larger; on the contrary, in the LTD regime, the relative G_{OFF} reduction is larger when the initial G_{OFF} is larger. The curves corresponding to the LTP and LTD regimes are fitted to exponential and sigmoid functions, respectively, for different programming times, as shown in Figure 38(b). As expected, a longer programming time consistently leads to a larger G_{OFF} change because of the larger V_T change caused by more trapped/de-trapped charge [22].

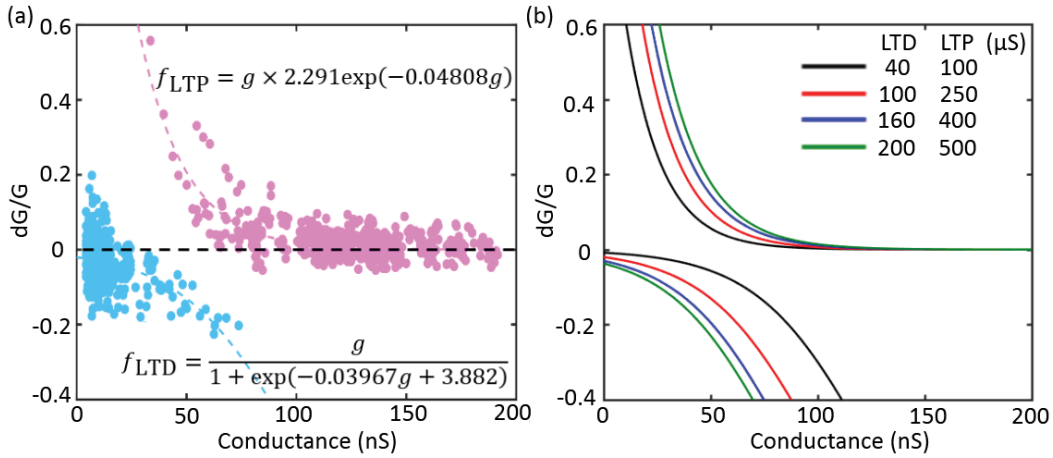


Figure 38: (a) Weight-Dependent Plasticity when Five Trapping/Detrapping Pulses are applied in the LTD/LTP Regimes, respectively and (b) Fitted Curves when Pulses of different Widths are applied

2.3 Use of CTT-based Analog Memory in Unsupervised Learning Systems

2.3.1 Winner-Takes-All Clustering Network

Owing to the characteristics described above, CTTs can be used as synapse devices in neural networks. As a demonstration vehicle, we describe here a one-layer winner-takes-all (WTA) neural network aiming at clustering stylized letters z, v, n, and one-bit-flipped noisy versions of them (Figure 39(a)) [23]. The input layer of the network has nine neurons corresponding to nine pixels of the pattern and the output layer has three neurons corresponding to the three categories: z, v, and n, respectively (Figure 39(b)). For each output neuron j (1, 2, or 3), the output is determined by $y_j = \sum_{i=1}^{9} x_i G_{OFFi,j}$, where $G_{OFFi,j}$ is the G_{OFF} of the CTT between the input neuron i and the output neuron j, and x_i is the input which is 50 mV when the ith pixel is black (firing) or 0 when the i-th pixel is white (not firing). For each presentation of a pattern, the neuron with the largest output fires, and the 9 synaptic weights associated with this neuron are updated with a WTA rule, such that a synapse is strengthened if the input neuron also fires, or weakened if the input neuron does not fire. If a synapse is strengthened, the CTT is programmed by a de-trapping

pulse; if the synapse is weakened, the CTT is programmed by a trapping pulse. In the simulation, we start from CTTs with random G_{OFF} ranging from 50−150 nS. Training of the neural network starts with randomly selecting a pattern from z, v, or n, and presenting it to the network. Then a random bit of the pattern is flipped and the noisy version is presented to the network again. Expressions fitted from experimental data are used to update the synaptic weights. The entire process is free of any intervention.

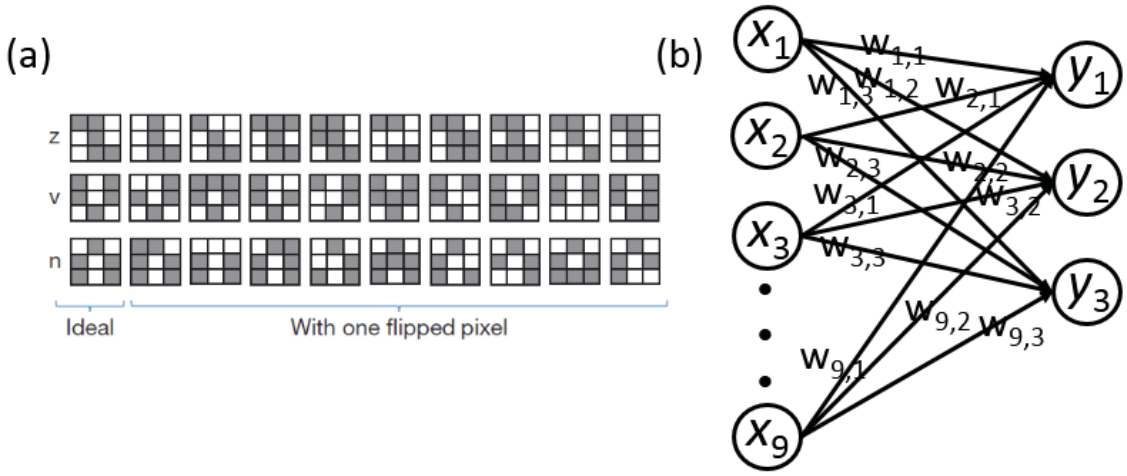


Figure 39: (a) Stylized Letters z, v, n, and One-Bit-Flipped Noisy Versions of them (adapted from [23]) and (b) Setup of the Unsupervised Neural Network

In the simulation, a total of 1,000 patterns are presented to the neural network with 500 correct patterns and 500 noisy patterns. Two trapping and de-trapping pulses as specified above are applied during the LTD and LTP regimes. The clustering results for the first and the last 100 presentations are, respectively, shown in Figure 40(a) and Figure 40(b). It is observed that a substantial number of misclassifications occur in the first 100 cycles, while all patterns are correctly classified for the last 100 cycles. To better understand the convergence behavior of the algorithm, a specialization function, S_i, is defined for each output neuron i, as the pattern **x** (z, v, or n) which yields the largest output y_i for the neuron. Perfect clustering is achieved when the neuron specializations remain constant and correspond to three different patterns as the neural network is trained. The specializations of the output neurons as the network is trained are shown in Figure 40(c). In fact, perfect clustering is achieved after only 82 training cycles after which Neurons 1, 2, and 3 correspond to patterns n, v, z, respectively. It should be further noted that this example is only to illustrate the evolution of specializations and does not represent a typical case. It is verified through 10,000 simulation runs that, the average number of cycles after which perfect clustering is achieved is only 24, well within the demonstrated endurance of over 1,000 for CTT-based non-volatile memory [22].

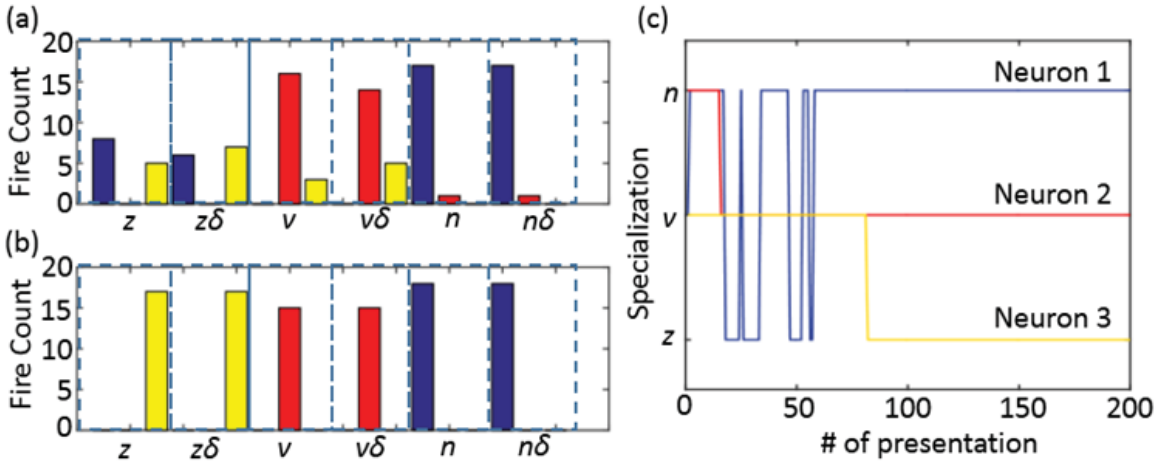


Figure 40: Fire Counts from Three Output Neurons (a) before and (b) after Training, and (c) Evolution of the Output Neuron Specializations as the Network is trained

For (a) and (b) blue, red, yellow: output neurons 1, 2, and 3. "δ" denotes a noisy version.

An example of the evolution of the synaptic weights $G_{OFF1,1}$ and $G_{OFF2,1}$is depicted in Figure 41. It is observed that, the sharp decreases in $G_{OFF2,1}$ are larger than the sharp increases in $G_{OFF1,1}$, which is caused by the asymmetry between LTP and LTD found in Fig. 37(c). It is also observed that, the weights, starting from random values, eventually reach a steady state after which each weight only varies around a certain value. In this example, the steady-state is 23.8 nS for $G_{OFF1,1}$ and 93.2 nS for $G_{OFF2,1}$ for the last 100 cycles when two trapping/de-trapping pulses are applied in the LTD/LTP regimes. These two values, representing respectively "low" and "high" weights after training, vary with the applied programming conditions. For instance, when five trapping/de-trapping pulses are applied, a "low" of 15.2 nS and a "high" of 95.8 nS are obtained. When a longer programming pulse is applied, larger G_{OFF} change is induced in each update step, leading to higher "high" and lower "low" eventual weights. Larger weight changes also result in faster convergence and a smaller noise margin. It is anticipated that the amplitudes of the trapping/de-trapping pulses will have similar effects on the convergence behavior.

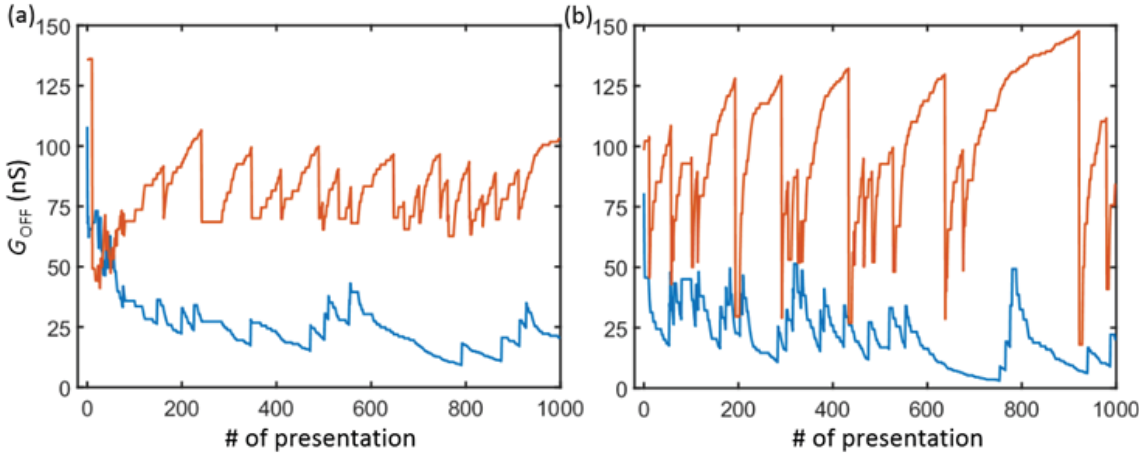


Figure 41: Example of the Evolution of Synaptic Weights $G_{OFF1,1}$ (blue) and $G_{OFF2,1}$ (red) for different Programming Times

(a) Two pulses are applied for LTD/LTP and (b) five pulses are applied for LTD/LTP

2.3.2 Robustness to Variation

In practice, when actual CTTs are used to construct the neural networks, the effect of device variation on the robustness of the algorithm needs to be evaluated. We illustrate here the example where two trapping and de-trapping pulses are used to update the weights (shown in Figure 42(a)). An empirically determined variation of Gaussian distribution with 3σ of $f_{10pulse} - f_{2pulse}$ is added to the conductance change calculated from fitted expressions, where $f_{10pulse}$ denotes the fitted conductance change when ten pulses are used to update the weights and f_{2pulse} denotes the fitted conductance change when two pulses are used to update the weights. More variation is introduced when $G_{OFF} > 60$ nS in the LTP regime and when $G_{OFF} < 40$ nS in the LTD regime to better approximate the experimental data. With this variation, the simulation was performed for 10,000 times and a 100% perfect clustering rate was achieved. An example of ΔG_{OFF} as a function of G_{OFF} from one of these simulations is depicted in Figure 42(b). It is indeed observed that the conductance change with the empirically introduced variation is comparable to the experimental data. With this methodology, it is also found that a longer programming time leads to a less robust algorithm: perfect clustering cannot be achieved when five LTP/LTD pulses are applied. It means that the effects of the variation are smaller when the programming time is shorter. This is because a shorter programming time corresponds to a smaller ΔG_{OFF} in each update step.

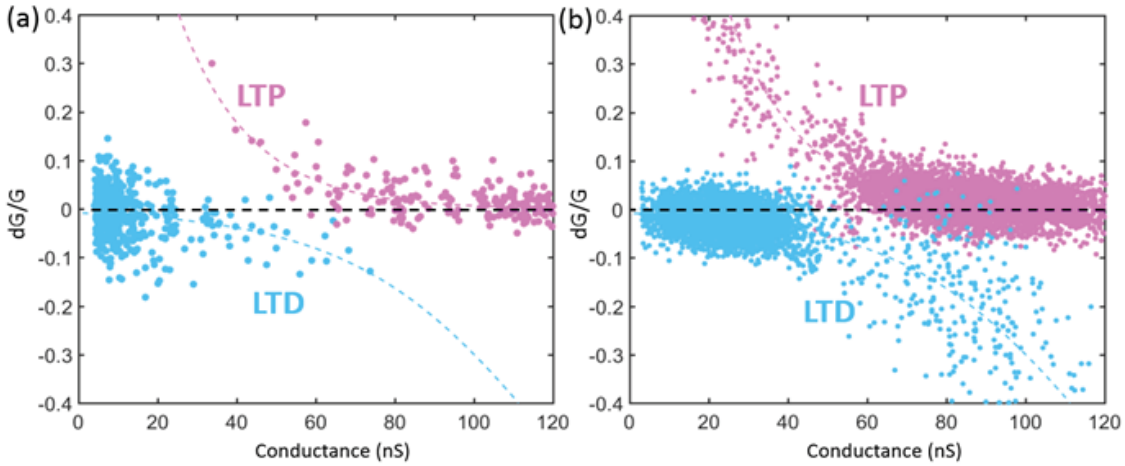


Figure 42: (a) Experimentally measured and (b) Empirically determined Relative Conductance Change as a Function of the Conductance itself in the LTP and LTD Regimes
The algorithm converges with the variation shown in Figure (b).

2.4 Use of CTT-based Analog Memory in Supervised Learning Systems

During the course of this study, we have also explored the use of CTT arrays to implement a critical function in any neural works: matrix multiplication to compute the summation of weighted inputs (shown in Figure 43(a)) [24]. This function is important in both back-propagation training process and feed-forward inference mode.

As in the case of unsupervised learning, the channel conductance (G_{ch}) of the CTT, at a given V_G, is used as the synaptic weight. As we have shown earlier, G_{ch} can be continuously modulated by very short (~20 μs) gate pulses, making the weight-tuning very accurate. Conceptually, how a

CTT array is used to calculate the weighted sum is depicted in Figure 43(b). The drain voltages ($V_i^{(0)}$) proportional to the pixel intensities of a pattern are applied to the CTTs and weighted by G_{ch} as the channel currents, which are then added together according to Kirchhoff's current law to be fed to the op-amp. Passed through an activation function (can be implemented by a differential pair or a linear rectifying unit), these weighted sums are then used as the inputs ($V_i^{(1)}$) to the next layer, and the same process is performed again.

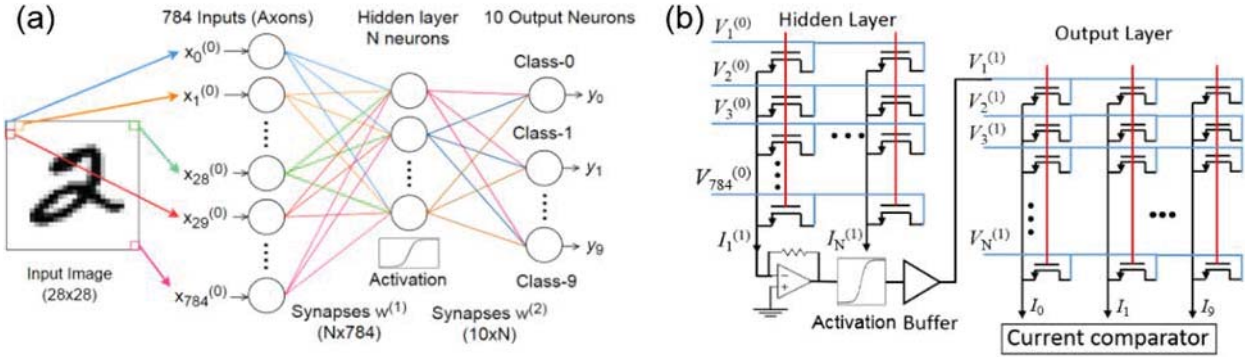


Figure 43: (a) Typical fully connected Neural Network and (b) Hardware Schematics of CTT Arrays to implement such Neural Networks

To demonstrate the potential of CTT arrays for neuromorphic computing, a classic problem – handwritten digit recognition using the MNIST (Modified National Institute of Standards and Technology) dataset – was studied with CTTs as the synapses. We first trained a two-layer perceptron network using batch-mode delta rule with back-propagation [24]. This CTT-based network contains 784 inputs (28×28 images) and 10 outputs (0−9) and a 97% recognition accuracy is obtained. This accuracy is comparable with the highest reported using a two-layer perceptron network. The trained weights were then transformed into CTT-compatible, nonnegative values corresponding to achievable CTT channel conductance values − determined to be 10−180 nS from experimental data [25] − using a linear one-to-one mapping. These weights are next loaded into an array of CTT devices by applying trapping/de-trapping pulses separately to each device, altering their respective threshold voltage. By changing the threshold of each device, we can program an array of channel conductance values. The programmed array can then be used for matrix multiplication.

2.4.1 CTT-based Inference Engine

Based on this fundamental understanding, a CTT-based inference engine can be built.

(i) System Architecture

The top-level block diagram of the system is depicted in Figure 44.

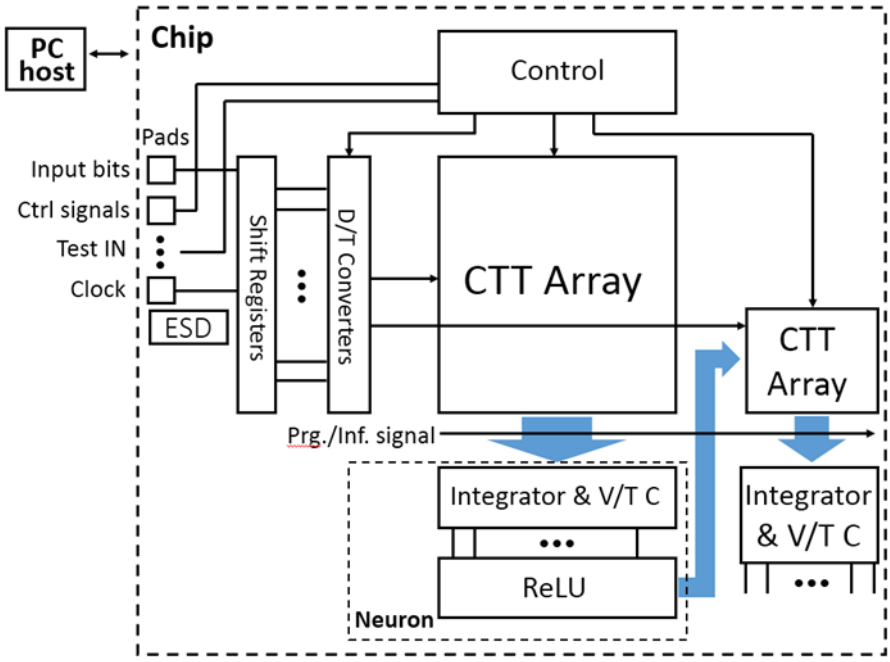


Figure 44: Block Diagram showing the System Architecture

The system works in two phases: programming and inference. During the programming mode, the CTTs in the array are programmed to desired conductance according to a software-trained model. During the inference mode, a pulse-width modulation (PWM) scheme is employed. Pulse-width, instead of voltage amplitude, is used here to carry information because we want to avoid the expensive (area and power) digital-to-analog converters. The pixel intensities of an image are first loaded onto the chip and stored in the shift registers. These inputs are next converted (through digital-to-time converters) to pulses whose width are proportional to the pixel intensities. The pulses are fed to the drains of the CTT devices, which, with a proper gate bias, generate current pulses whose amplitudes are determined by the channel conductance (weights) and whose widths are determined by the drain pulses (inputs). The current pulses for a particular neuron are integrated on a capacitor, generating a voltage that is proportional to the weighted sum of all the inputs. The signals to the next layer are also pulse-width modulated. This requires the linear discharge of the capacitor, which can be done using a current source. After the output PWM signals are generated, they are sent to the rectifying linear unit (activation function) before being sent to the next layer as inputs. For a multi-layer neural network, the same process can be repeated until the last layer, which generates a number of PWM signals. The system then detects which signal has the longest duration and determines the corresponding classification.

(ii) Twin-Cell Architecture

To implement the synapse, a twin-cell architecture, instead of a single cell, is used (Figure 45). The weight w of the synapse is represented by $w^{+} - w^{-}$.

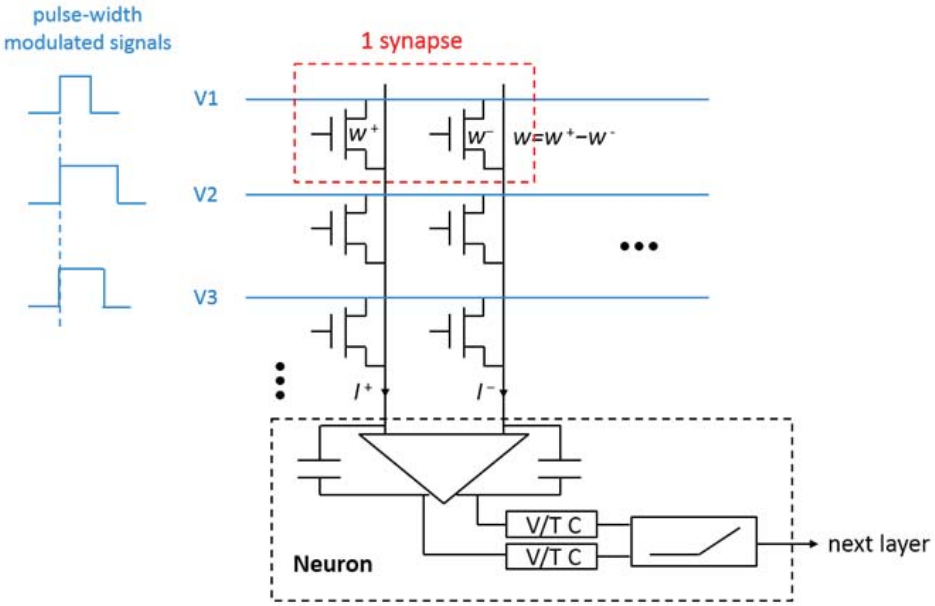


Figure 45: Twin-Cell Synapse Architecture

We adopt this design because: (1) it avoids the conversion from bipolar mathematical weights to unipolar channel conductance; by taking the difference between w^+ and w^-, the physically implemented weights are bipolar. (2) It reduces the effects of variation since the virgin devices will be very close to each other and similar.

The synapses corresponding to a neuron are implemented by CTTs in two columns. The drains of the CTTs in different rows are different PWM signals carrying input information while the sources of the CTTs are tied and fed to the integrator.

(iii) Programming Scheme

Before the CTT array can be used for a matrix multiplication engine, the conductance of the CTTs first need to be programmed to proper levels. This is enabled by adding a programming switch in the end of each column. The array is programmed column-by-column, as shown in Figure 46. When a particular column is being programmed, its programming switch is turned on and the integrator is bypassed. The DMUX selects this column and passes the programming V_G pulse to the gates of the CTTs in this column. Drain pulses (V_1−V_N) of different widths are sequentially applied to program the CTTs to desired conductance. Once a column is done programming, the programming switch of the next column is turned on and the CTTs in that column are programmed.

The reason why we need to program the CTTs individually, instead of applying drain pulses to all CTTs in a column, is that, the CTT programming current is very high (~1.5 mA). If all CTTs in a column are programmed at the same time, the programming switch needs to be very large in order to carry that large current, adding a significant capacitance to the integrator and complicating its design substantially. By programming the CTTs individually, the size of the programming switch is only approximately three times that of a CTT. The price we have to pay with individual programming is increased programming time. For example, for a hidden layer of

size 800×100, the programming time will be 13.3 minutes assuming a 10 ms individual programming time.

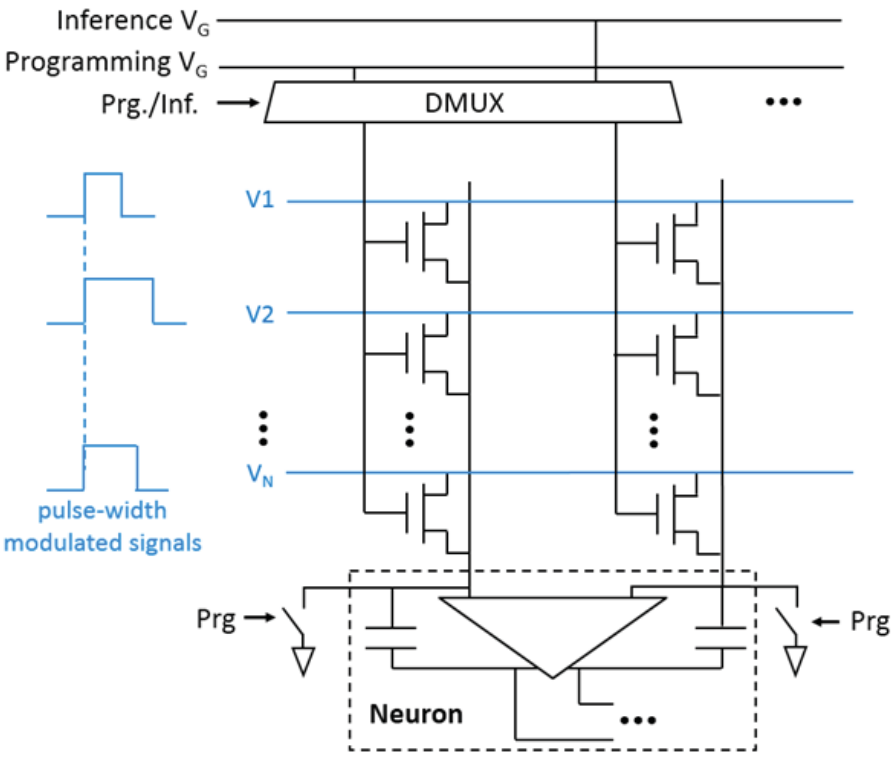


Figure 46: Array Programming Scheme

(iv) Neuron Design

Up until now, we have assumed a fully differential op amp for the integrator. There is an alternative approach to it (shown in Figure 47). During the integration phase of the inference mode, the current source is off and the incoming current is from the CTTs in the column. This incoming current charges the capacitor C_{INT} and increases the voltage across it. V_{ref} is chosen to be high enough such that the PFET is always in saturation mode, even though the voltage across the capacitor increases. In the end of the integration phase, the voltage across C_{INT} represents the weighted sum, as we have discussed earlier. Then this neuron starts the output phase. The voltage across C_{INT} is discharged through the current source, until it is below a certain threshold when the output turns 0. This low output turns off the current source and ends the output phase.

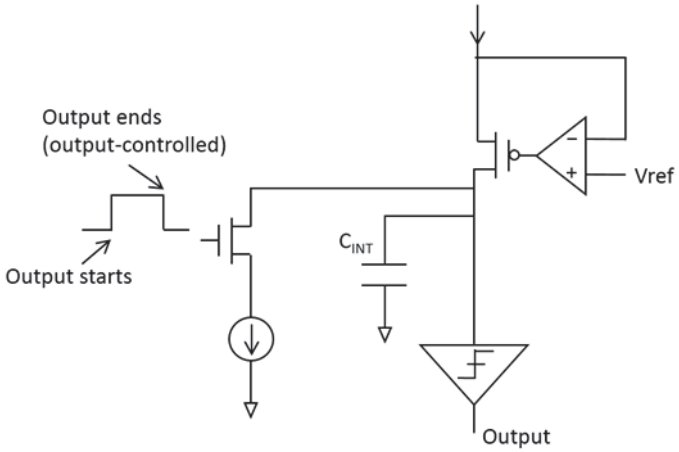


Figure 47: Neuron Design

(v) Rectifying Linear Unit

The rectifying linear unit (ReLU) (shown in Figure 48) is adopted as the activation function primarily because the nonnegative output eliminates the conversion from bipolar outputs (such as those from a sigmoid) to nonnegative pulse widths. Interestingly, this choice also presents us with an extra bonus – a very simple hardware implementation. Logically, after the currents in the two CTT columns are integrated as the voltages V^+ and V^- across the capacitor, the difference, $V^+ - V^-$, needs to be first computed by additional circuitry and then converted to pulse-width-modulated signals. However, in the context of pulse-width-modulation, this can be conveniently implemented using a simple logic gate. Consider Figure 49(a) and Figure 49(b), where two different cases were shown: $V^+ > V^-$ and $V^+ < V^-$, respectively. In Figure 49(a) where $V^+ > V^-$, the output is a pulse which is only nonzero when the V^+ pulse is high; on the hand, when $V^+ < V^-$ as shown in Figure 49(b), the output is zero. This function can be described by the truth table in Figure 49(c). Therefore, the function can be realized by the simple logic circuit shown in Figure 49(d).

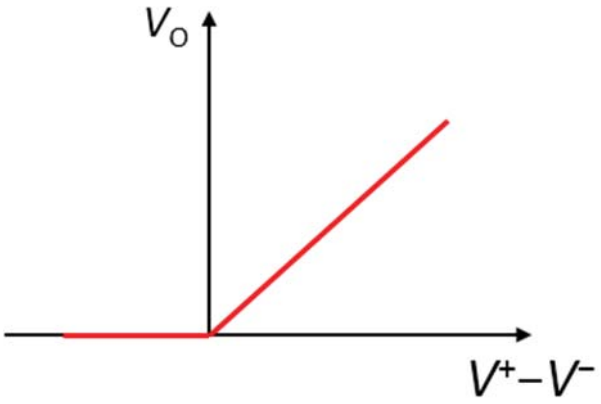
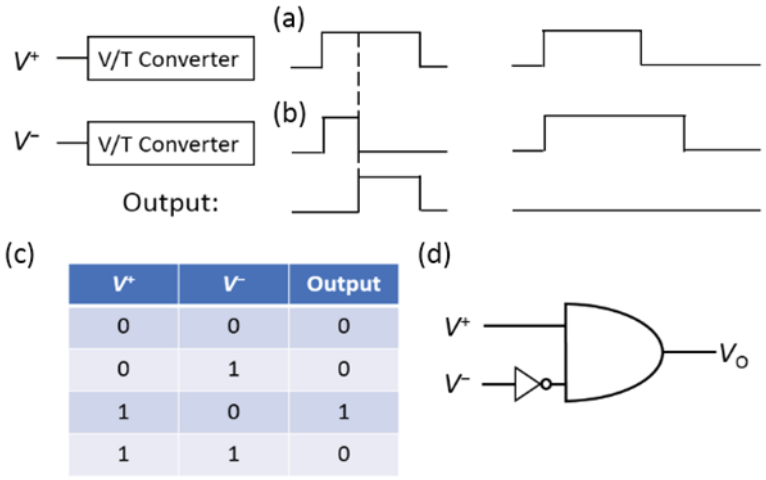


Figure 48: ReLU Activation Function



V+	V−	Output
0	0	0
0	1	0
1	0	1
1	1	0

Figure 49: (a) ReLU in the Context of PWM Signals and (b) Implementation of ReLU with Simple Logic Circuit

2.4.2 Considerations of Imperfections

(i) Impact of Weight Precision on Classification Performance

In a real hardware system, one does not have access to all possible analog weights; also, the op-amp cannot differentiate between an infinite number of total current levels. In this study, we investigated how many levels of weights and hidden layer currents are necessary for the CTT-based network to have an acceptable accuracy.

The result is shown in Figure 50. It can be observed that, when there are 5 or more bits in the hidden layer current (which is not a stringent requirement on the op-amp), 6 bits in the weights is enough. In fact, the improvement in accuracy when bits of weights is 7, compared to when it is 6, is only 0.2%. Considering that we have been able to achieve over 300 mV in V_T and sub-5 mV resolution, 6 bits of weight accuracy is achievable in actual hardware.

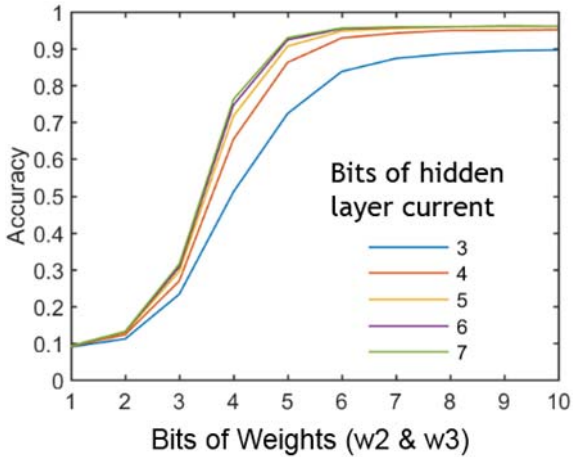


Figure 50: Accuracy vs. Bits of Weights (w2 and w3) for different Bits of Hidden Layer Current

(ii) Impact of Noisy Weights on Classification Performance

One inherent property of any neuromorphic system, just as in human brains, is its fault-tolerance. In a CTT-based neuromorphic system, random circuit noises are anticipated. However, from our studies, we find that a CTT-based system is very robust to these noises. For each testing pattern, a Gaussian noise of a standard deviation σ is added to the weighted sum, and the recognition accuracy is evaluated. For each σ value, the simulations were run 100 times, and the box plot is shown in Figure 51. It can be observed that when σ = 10 nA (2% of the maximum hidden layer current as shown in the inset of Figure 51), the average degradation is only 0.3%. Although the accuracy degrades fairly fast as σ increases, a degradation of less than 1% is anticipated when σ is within 4% of the maximum hidden layer current. This should not impose a stringent requirement on the noise level in the circuits.

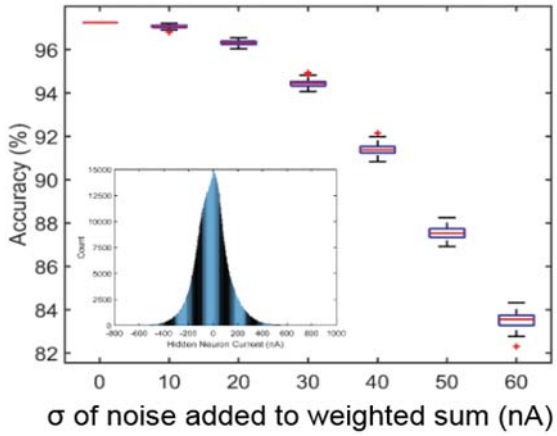


Figure 51: Box Plot of Recognition Accuracy vs. σ of the Added Noise
The inset shows a histogram of hidden layer current.

2.5 Summary and Future Work

Besides studying the CTT physics for multi-time programmable digital memory, we also investigated the potential of CTT for analog memory applications. Specifically, experimental data from 22-nm SOI devices reveal that a charge-trap transistor possesses promising characteristics for implementing synapses in neural networks, such as spike-timing dependent plasticity, very fine tunability, weight-dependent plasticity, and low power consumption. A proof-of-concept winner-takes-all neural network is simulated based on experimental data and perfect clustering is achieved within tens of training cycles. This means that the network can be trained for multiple times, and a larger system can be built. The robustness of the procedure to the device variation is also studied.

We are also working on demonstrating a supervised learning neural network featuring CTTs as the synapses. We have decided to use PWM, twin-cell synapse design, and ReLU activation function. The system is found to be very robust to device variations and circuit noises. The weight accuracy requirement is not very stringent – it is found that 6 bits of CTT weights (achievable from our earlier experiments) can already provide a good enough classification accuracy.

Ongoing efforts include the design and tapeout of a CTT-based neuromorphic chip for digit recognition, and more elaborate designs that address programming time, scalability, power/area reduction, redundancy, unsupervised learning, etc. in the long-term.

3. References

[1] E. Cartier, B. P. Linder, V. Narayanan, and V. K. Paruchuri, "Fundamental understanding and optimization of PBTI in nFETs with SiO_2/HfO_2 gate stack", IEDM, 2006.

[2] C. Kothandaraman, X. Chen, D. Moy, D. Lea, S. Rosenblatt, F. Khan, D. Leu, T. Kirihata, D. Ioannou, G. LaRosa, J. B. Johnson, N. Robson, and S. S. Iyer, "Oxygen vacancy traps in Hi-K/Metal gate technologies and their potential for embedded memory applications", IRPS, 2015.

[3] H. Hamamura, T. Ishida, T. Mine, Y. Okuyama, D. Hisamoto, Y. Shimamoto, S. Kimura, and K. Torii, "Electron trapping characteristics and scalability of HfO_2 as a trapping layer in SONOS-type flash memories", IRPS, 2008.

[4] E Cartier and A Kerber, "Stress-Induced Leakage Current and Defect Generation in nFETs with HfO_2/TiN Gate Stacks during Positive-Bias Temperature Stress", IRPS, 2009.

[5] F. Crupi, R. Degraeve, A. Kerber, D.H. Kwak, and G. Groeseneken, "Correlation between Stress-Induced Leakage Current (SILC) and the HfO_2 Bulk Trap Density in a SiO_2 / HfO_2 Stack", IRPS, 2004.

[6] S. Narasimha, P. Chang, C. Ortolland, D. Fried, E. Engbrecht, K. Nummy, P. Parries, T. Ando, M. Aquilino, N. Arnold, R. Bolam, J. Cai, M. Chudzik, B. Cipriany, G. Costrini, M. Dai, J. Dechene, C. DeWan, B. Engel, M. Gribelyuk, D. Guo, G. Han, N. Habib, J. Holt, D. Ioannou, B. Jagannathan, D. Jaeger, J. Johnson, W. Kong, J. Koshy, R. Krishnan, A. Kumar, M. Kumar, J. Lee, X. Li, C-H. Lin, B. Linder, S. Lucarini, N. Lustig, P. McLaughlin, K. Onishi, V. Ontalus, R. Robison, C. Sheraw, M. Stoker, A. Thomas, G. Wang, R. Wise, L. Zhuang, G. Freeman, J. Gill, E. Maciejewski, R. Malik, J. Norum, and P. Agnello, "22nm High-Performance SOI Technology Featuring Dual-Embedded Stressors, Epi-Plate High-K Deep-Trench Embedded DRAM and Self-Aligned Via 15LM BEOL", IEDM, pp. 52-55, 2012.

[7] A. Kerber, S. Krishnan, and E. Cartier, "Voltage Ramp Stress for Bias Temperature Instability Testing of Metal-Gate/High-k Stacks", IEEE Electron Device Letters, 30 (12), 2009.

[8] P. Su, K. Goto, T. Sugii, and C. Hu, "Excess Hot-Carrier Currents in SOI MOSFETs and Its Implications", IRPS, 2002.

[9] D. Dallmann and K. Shenai, "Scaling Constraints Imposed by Self-Heating in Submicron SOI MOSFET's", Transactions on Electron Devices, 42 (3), 1995.

[10] N. Raghavan, K. Pey, and K. Shubhakar. "High-κ dielectric breakdown in nanoscale logic devices - Scientific insight and technology impact", Microelectronics Reliability, 54 (5), pp. 847-860, 2014.

[11] G. Bersuker, J. Sim, C. Park, C. Young, S. Nadkarni, R. Choi, and B. Lee, "Mechanism of Electron Trapping and Characteristics of Traps in HfO_2 Gate Stacks", IEEE Transactions on Device and Materials Reliability, 7 (1), 2007.

[12] T. Grasser, P.-J. Wagner, H. Reisinger, Th. Aichinger, G. Pobegen, M. Nelhiebel, and B. Kaczer, "Analytic modeling of the bias temperature instability using capture/emission time maps", IEDM, 2011.

[13] Y-P. Gong, A-D. Li, X. Qian, C. Zhao, and D. Wu, "Interfacial structure and electrical properties of ultrathin HfO_2 dielectric films on Si substrates by surface sol–gel method", J. Phys. D: Appl. Phys., 42 (2009).

[14] E. P. Gusev and C. P. D'Emic, "Charge detrapping in HfO2 high-κ gate dielectric stacks", Applied Physics Letters, 83 (25), 2003, pp.5223-5225.
[15] J. Viraraghavan, D. Leu, B. Jayaraman, A. Cestero, R. Kilker, M. Yin, J. Golz, R. R. Tummuru, R. Raghavan, D. Moy, T. Kempanna, F. Khan, T. Kirihata, and S. S. Iyer, "80Kb 10ns Read Cycle Logic Embedded High-K Charge Trap Multi-Time-Programmable Memory Scalable to 14nm FIN with no Added Process Complexity", Symp. VLSI-Circuits, June 2016, pp. 1-2.
[16] B. Jayaraman, D. Leu, J. Viraraghavan, A. Cestero, M. Yin, J. Golz, R. R. Tummuru, R. Raghavan, D. Moy, T. Kempanna, F. Khan, T. Kirihata, and S. S. Iyer "80-kb Logic Embedded High-K Charge Trap Transistor-Based Multi-Time-Programmable Memory With No Added Process Complexity", JSSC, January 2018.
[17] T. Hook, F. Allibert; K. Balakrishnan, B. Doris, D. Guo; N. Mavilla, E. Nowak, G. Tsutsui, R. Southwick, J. Strane, X. Sun, "SOI FinFET versus bulk FinFET for 10nm and below", IEEE S3S Conf., 2014, pp. 1-3.
[18] D. Jang, E. Bury, R. Ritzenthaler, M. Garcia Bardon, T. Chiarella, K. Miyaguchi, P. Raghavan, A. Mocuta, G. Groeseneken, A. Mercha, D. Verkest, and A. Thean, "Self-heating on bulk FinFET from 14nm down to 7nm node", IEEE IEDM, 2015, pp. 11-6.
[19] C. Kothandaraman, S. K. Iyer, and S. S. Iyer, "Electrically Programmable Fuse (eFUSE) Using Electromigration in Silicides", IEEE Electron Device Letters, 23 (9), 2002.
[20] G.-Q. Bi and M.-M. Poo, "Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type," J. Neurosci., vol. 18, no. 24, pp. 10 464–10 472, Dec. 1998.
[21] D. Kuzum, S. Yu, and H. S. P. Wong, "Synaptic Electronics: Materials, Devices, and Applications," Nanotechnology, vol. 24, 382001.1-22, Sept. 2013. doi: 10.1088/0957-4484/24/38/382001
[22] F. Khan, E. Cartier, J. C. S. Woo and S. S. Iyer, "Charge Trap Transistor (CTT): An Embedded Fully Logic-Compatible Multiple-Time Programmable Non-Volatile Memory Element for High-k-Metal-Gate CMOS Technologies," IEEE Electron Device Letters, vol. 38, no. 1, pp. 44-47, Jan. 2017. doi: 10.1109/LED.2016.2633490
[23] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and Operation of An Integrated Neuromorphic Network Based on Metal-oxide Memristors," Nature, vol. 521, pp. 61-64, May 2015. doi: 10.1038/nature14441
[24] J. Hertz et al., "Introduction to the Theory of Neural Computation," (Perseus, 1991).
[25] X. Gu and S. S. Iyer, "Charge-trap transistors for unsupervised learning," IEEE Electron Device Lett., IEEE Electron Device Letters 38 (9), 1204-1207.

List of Abbreviations, Acronyms, and Symbols

ACRONYM	DESCRIPTION
CMOS	Complementary Metal-Oxide-Semiconductor
CTT	Charge Trap Transistor
DARPA	Defense Advanced Research Projects Agency
ERS	Erase
FDSOI	Fully Depleted Silicon On Insulator
FET	Field-Effect Transistor
FN	Fowler–Nordheim
G_{ch}	Channel Conductance
GIDL	Gate Induced Drain Leakage
G_{OFF}	OFF-conductance
HKMG	High-K-Metal-Gate
IFL	Interfacial Layer
I_g	Gate Injection
IL	InterLayer
L	Channel Length
LTD	Long-Term Depression
LTP	Long-Term Potentiation
MOS	Metal-Oxide-Semiconductor
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MTPM	Multi-Time Programmable Memory
NVM	Non-Volatile Memory
OTPM	One Time Programmable Memory
P/E	Program/Erase
PDA	Post Deposition Anneal
PDSOI	Partially Depleted Silicon On Insulator
PF	Poole-Frenkel
PRG	Program
PVRS	Pulsed Gate Voltage Ramp Sweep
PWM	Pulse-Width Modulation
ReLU	Rectifying Linear Unit
R_{th}	Thermal Resistance
SOI	Silicon On Insulator
SR	Schottky-Richardson
STDP	Spike-Timing Dependent Plasticity
TAT	Trap-Assisted Tunneling
TDDB	Time Dependent Dielectric Breakdown
V_d	Drain Bias
V_g	Gate bias
V_{gd}	Gate-to-Drain Bias
V_T	Threshold Voltage
W_{ch}	Channel Width
WTA	Winner-Takes-All
ΔV_T	Threshold Voltage Shift